DOCUMENT RESUME

ED 302 700 CE 051 729

TITLE Juvenile Counselor Validation Report and Technical Adequacy Report. Standards and Training for Corrections Program.
INSTITUTION California State Board of Corrections, Sacramento.
PUB DATE 87
NOTE 133p.; For related documents, see CE 051 717-725, 730-735, and 739-744.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Correctional Rehabilitation; *Counselor Evaluation; Counselor Qualifications; Counselors; *Counselor Selection; Counselor Training; Delinquency; Employment Qualifications; Entry Workers; Evaluation Criteria; *Job Analysis; *Occupational Tests; Postsecondary Education; Predictive Validity; Research Methodology; *Test Validity
IDENTIFIERS *California

ABSTRACT

These two reports describe the research, analysis, and results of the second phase of a three-phase research study sponsored by the Standards and Training for Corrections Program (STC) of the California Board of Corrections. The validation report consists of seven chapters. Following an introductory chapter that explains the rationale for the program and documents the approach, the remainder of the report describes the methods used to develop and validate entry-level selection procedures, especially for juvenile counselors. Chapter 2 summarizes the job analysis results that formed the foundation for predictor and criterion development. Development of the predictors and criteria are described in Chapters 3 and 4. Chapter 5 describes the methods used to collect predictor and criterion data from a sample of approximately 200 juvenile counselor incumbents and their supervisors. Chapter 6 reports the methods used to analyze the predictor and criterion data and documents the results of those analyses. Finally, Chapter 7 reviews the results with an eye toward their interpretation and the implications they have for use of the Juvenile Counselor examination as part of the preemployment selection process of participating agencies. The technical adequacy report examines the criterion data collection and the adequacy of the sample to confirm the technical adequacy of the research described in the validation report. (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302700

JUVENILE COUNSELOR VALIDATION REPORT AND TECHNICAL ADEQUACY REPORT

STANDARDS AND TRAINING FOR CORRECTIONS PROGRAM

CE051729

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

STANDARDS AND TRAINING
FOR CORRECTIONS PROGRAM

STC

BOARD OF CORRECTIONS

VALIDATION REPORT

JUVENILE COUNSELOR

Copyright 1987 California Board of Corrections

In preparing the Juvenile Counselor Examination for the California Board of Corrections, Personnel Decisions, Inc. has taken into account the following documents concerning selection testing:

- The American Psychological Association, Standards for Educational and Psychological Tests, (1974).
- The APA Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures, 2nd edition, (1980).
- The Equal Employment Opportunity Commission, Uniform Guidelines on Employee Selection Procedures, (1978).

CONTENTS

		Page
	List of Tables	ii
	List of Figures	iv
Chapter 1	INTRODUCTION	
	Project Background	1
	Participating Departments	2
	Research Strategy and Rationale	2
	Report Overview	4
Chapter 2	SUMMARY OF THE JOB ANALYSIS	5
Chapter 3	PREDICTOR DEVELOPMENT	
	Overview and Rationale	7
	Linking KSAs to Job Tasks	9
	Grouping Task Items into Task Categories	13
	Task Category to Worker Characteristic Ratings	13
	Setting Linkage Criteria	19
	Linking Tests to Worker Characteristics	19
	Linking Tests to Task Categories	31
	Predictor Development	35
	Test construction	35
	Pretesting the preliminary tests	39
	Item analysis and revision	41
	Special revision methods	43
	Final Test Battery	44
	Noncognitive Tests	44
	Sample Items Booklet	45
Chapter 4	CRITERION DEVELOPMENT	
	Collecting Critical Incidents	46
	Revising and Retranslating the Critical Incidents	47
	Developing Behavioral Summary Scales	52
	Other Ratings Scales	52
	Worker characteristic ratings	52
	Overall effectiveness	54
Chapter 5	DATA COLLECTION	
	Predictor Data Collection	55
	Criterion Data Collection	56
	Final Validation Sample	57
	Missing name	57

CONTENTS (Chapter 5 continued)

		Page
	Missing ratings	59
	Low tenure	59
	Inaccurate ratings	59
	Demographics of Final Validation Sample	59
	Technical Adequacy of Sample	63
Chapter 6	ANALYSES AND RESULTS	
	Criterion Analyses	65
	Data collection and cleaning	65
	Developing an overall criterion	67
	Search for additional criteria	70
	Criterion descriptive statistics	73
	Predictor Analyses	78
	Data collection and cleaning	78
	Scale revisions and final psychometrics	78
	Descriptive statistics	79
	Validity and Fairness Analyses	85
	Adverse impact and test fairness	85
	Validity and fairness of the individual tests	86
	Developing a Final Selection Battery	88
	Differential predictability of specific criteria	93
Chapter 7	DISCUSSION AND RECOMMENDATIONS	
	Exam Format and Content	95
	Using the Juvenile Counselor Examination	99
	Exam Security	99
	Scoring and Data Collection	100
	Score Interpretation and Cutoff Scores	102
References		103

TABLES

		Page
3.1	Definitions of 26 Juvenile Counselor Worker Characteristics	10-12
3.2	50 Task Categories for Juvenile Counselors	14-16
3.3	Average Importance Ratings of 26 Worker Characteristics for 50 Task Categories	17-18
3.4	Reliabilities and Ratings Summary for Juvenile Counselors	20
3.5	Average Usefulness Ratings of 72 Tests as Measures of 18 Before-Hire Worker Characteristics	22-30
3.6	Reliabilities and Ratings Summary of Test to Worker Characteristics Ratings	32
3.7	Linkage Criteria and Summary for Test to Worker Characteristic and Worker Characteristic to Task Category Ratings	33-34
3.8	Cognitive and Noncognitive Tests for the Experimental Selection Battery	36-37
3.9	Pretest Results and Item Statistics for the Experimental Selection Battery	40
4.1	Juvenile Counselor Performance Dimensions Used During Retranslation	49-51
5.1	Composition of the Final Validation Sample	58
5.2	Gender and Ethnic Status of Final Validation Sample	60
5.3	Education and Tenure of Final Validation Sample	61
5.4	Relationship of Incumbents to Supervisors in Final Validation Sample	62
6.1	Descriptive Statistics for 37 Job Performance Ratings Scales	66
6.2	Correlation of 37 Job Performance Rating Scales with the Overall Criterion	69

TABLES (continued)

		Page
6.3	Factor Descriptions of Five Specific Job Performance Criteria	71
6.4	Descriptive Statistics and Intercorrelations for Five Factor Score Criteria	74
6.5	Descriptive Statistics for the Overall and Specific Job Performance Criteria	75
6.6	Average Overall Criterion Scores by Supervisor and Incumbent Gender	76
6.7	Average Overall Criterion Scores by Supervisor and Incumbent Ethnic Status	78
6.8	Final Item Statistics for the Experimental Selection Battery	80
6.9	Descriptive Statistics on 15 Individual Tests for the Final Validation Sample	81
6.10	Intercorrelations Among Individual Tests	82
6.11	Descriptive Statistics on Individual Tests by Incumbent Gender	83
6.12	Descriptive Statistics on 15 Individual Tests by Incumbent Ethnic Status	84
6.13	Correlations of 15 Individual Tests with Overall Job Performance	87
6.14	Source of Items Selected to Augment the Final Selection Battery	91
6.15	Gender and Ethnic Group Statistics for Juvenile Counselor Examination Scores	92
6.16	Correlations of Individual and Final Predictors with the Five Factor Criterion Measures	94
7.1	Description of the Noncognitive Tests Contributing Items to the Juvenile Counselor Examination	97-98
7.2	Juvenile Counselor Examination Score Distribution	101

FIGURES

		Page
3.1	Hypothetical Example of the Relationships Among Tests, KSAs, and Job Tasks	8
3.2	Test Development Plan for the Numerical Ability Test	38
4.1	Sample Performance Incident Form	48
4.2	Sample Behavioral Summary Scale	53
6.1	Factor Loadings of 36 Ratings Scales on Five Specific Criteria Factors	72

CHAPTER ONE: INTRODUCTION

The research, analysis, and results described in this report represent the second phase of a three-phase research study sponsored by the Standards and Training for Corrections Program (STC) of the California Board of Corrections. This introduction provides general information about the project and describes the strategy used to conduct the research.

Project Background

The California Board of Corrections had been charged by the State Legislature to undertake several previous studies of the probation (1964) and correctional systems (1971 & 1979). These studies found widely varying hiring standards and training practices from agency to agency -- even where printed job descriptions were quite similar. Additionally, many agencies were finding themselves in court faced with the task of defending their hiring procedures while incurring the considerable expense associated with such litigation and facing substantial delays in filling vacancies.

To address these problems, the Board of Corrections made two recommendations to the Legislature. The Board recommended that:

1. A state agency assume responsibility for an analysis and synthesis of job titles and job descriptions and for validation studies of selection standards for corrections personnel.

2. A commission be established, modeled after POST, specifically for standards setting and training for corrections personnel.

In response to the Board's recommendations, the California Probation, Parole, and Correctional Association (CPPCA) proposed Senate Bill 924 which became effective July 1, 1980. SB 924 created the Standards and Training for Corrections Program (STC) to establish minimum standards for the selection and training of local corrections and probation personnel, to certify training courses, and to administer the corrections training fund to help agencies comply with the training standards. Funds for this program are derived from penalty assessments on specified criminal offenses. Participation by individual agencies in the STC program is voluntary. Only those agencies accepting state funds through STC are required to adhere to the selection and training standards.

Subsequently, Assembly Bill 3296 was passed. This bill mandated the Board of Corrections to contract (among other things) for research on validated selection standards for corrections and probation personnel. This mandate provided the impetus for the research efforts described in this report and in related documents.

Participating Departments

County sheriff's or probation departments participating in the STC program were eligible to participate in this project. The 37 eligible sheriff's and 52 eligible probation departments were invited to attend one of three briefings where the goals and procedures of the study were presented. Approximately 20 sheriff's and 21 probation departments attended these briefings.

Departments unable to send representatives to the briefings were mailed copies of the briefing materials and participation agreement forms. Initially, 23 sheriff's departments and 30 probation departments (representing 62% and 58% of eligible departments) elected to participate in the study. The number varied from time to time as departments withdrew or were added to the study.

Research Strategy and Rationale

The research program has consisted of three phases, each corresponding to one of three objectives of the Standards and Training for Corrections Program. These three objectives are:

Phase 1 -- a statewide job analysis of three entry-level positions in county agencies: corrections officer, probation officer, and juvenile counselor.

Phase 2 -- the development and validation of entry-level selection procedures for these positions.

Phase 3 -- the development of training standards and curricula for these positions.

This report focuses primarily upon the second phase of the research project, namely the development and validation of entry-level selection procedures. It is important to note that this report concerns the juvenile counselor position only. Separate reports describe the parallel efforts undertaken for the probation officer and corrections officer positions also included in this project. Of necessity, this report refers frequently to procedures and analyses conducted during the job analysis phase of the project but is not intended to be a complete description of that research which is described fully elsewhere.[1]

[1]Grabow, K.M., Sevy, B.A., Houston, J.S. Statewide Job Analysis of Three Entry-level Corrections Positions. Personnel Decisions, Inc., 1983. (Available from the Board of Corrections.)

In addition to this report, information about the methods and results of the phase two research for juvenile counselors may be found in the User's Manual as well as in the Candidate Information Booklet and, to a lesser degree, in the test booklet. To maintain test security, access to the test booklets is strictly controlled and closely monitored.

A concurrent, criterion-related strategy was used to guide the development and validation of selection procedures for the entry-level juvenile counselor position. The basic goal of the selections standards phase of the project was to devise a method of accurately predicting the job performance of entry-level juvenile counselors. The criterion-related approach to meeting this goal consisted of developing measures of ability and personal characteristics (predictors) which could be shown to be statistically and meaningfully related to job performance. The predictors must also have been fair and not impose heavy administrative burdens to implement. A concurrent study is one in which predictor and criterion data are collected from job incumbents.

Professional standards (APA, 1980) suggest that a criterion-related approach is preferable whenever it is feasible and provide guidelines for determining the feasibility of such an approach. The first and, for this study, most important guideline concerns the stability of the job. Criterion-related studies are not feasible whenever the position being studied is in a period of rapid flux. Professional standards are most directly concerned with the stability of a position over time but, due to the variety of departments providing incumbent juvenile counselors to the study, stability of the position across agencies must also be considered. With respect to stability across agencies, job analysis data were analyzed to determine the extent to which the juvenile counselor job (as described by juvenile counselor incumbents and supervisors) varied as a function of several parameters (county size, type of facility, job shift, etc.). The results of these analyses clearly demonstrated the equivalence of the job from department to department, thus indicating that the stability of the job across agencies cannot be questioned.

The stability of the juvenile counselor position over time cannot be as directly addressed, but it seems clear that the position is not in a "period of rapid evolution". Nevertheless, the data in this report reflect only a two-and-one-half year "snapshot" of the juvenile counselor position. Should the job undergo significant changes in the future, it will be incumbent upon users of the selection procedures developed in this project to ensure that the procedures remain valid, fair, and useful.

Other guidelines relevant to determining the feasibility of a criterion-related strategy concern the potential for developing a relevant, reliable measure of job performance (criterion); and for collecting predictor and criterion data from a sample of persons that is "reasonably representative of the populations of peoples and jobs to which the results are to be generalized" (APA, 1980).

It is important that this sample be large enough to allow the reliable detection of meaningful relationships between the predictors and criteria. The research described in this report fully met each of these guidelines. Therefore, a criterion-related approach to the development and validation of entry-level selection standards for juvenile counselors was completely feasible.

Report Overview

The remainder of this report describes the methods used to develop and validate entry-level selection procedures. This description begins in Chapter Two with a summary of the job analysis results that formed the foundation for predictor and criterion development. Development of the predictors and criteria are described in Chapters Three and Four. Chapter Five describes the methods used to collect predictor and criterion data from a sample of approximately 200 juvenile counselor incumbents and their supervisors. Chapter Six reports the methods used to analyze the predictor and criterion data and documents the results of those analyses. Finally, Chapter Seven reviews the results with an eye toward their interpretation and the implications they have for use of the Juvenile Counselor Examination as part of the pre-employment selection process of participating agencies.

CHAPTER TWO: SUMMARY OF THE JOB ANALYSIS

The job analysis of the juvenile counselor position began with reviews of job descriptions, training and orientation materials, and performance evaluation forms supplied by participating probation departments. Document review was followed by site visits to eight counties. During the site visits, detention facilities were toured and juvenile counselor incumbents and supervisors interviewed individually and in groups to determine the major duties and tasks involved in the juvenile counselor position and the important requirements of the job.

Information gathered from the document review and site visits was used to develop a preliminary version of a job analysis questionnaire (JAQ). The preliminary questionnaire was reviewed by STC staff and by supervisors, training officers, and others from a sample of participating departments. Based on the reviewers' comments, the questionnaire was revised and then pilot tested on a sample of juvenile counselor incumbents and supervisors. During the pilot test, the incumbents and supervisors completed the preliminary questionnaire and provided feedback about its clarity and comprehensiveness.

Based upon the data collected in the pilot test, final revisions of the job analysis questionnaire were made. Two forms of this questionnaire were developed, one for juvenile counselor incumbents and the other for their supervisors. Each questionnaire consisted of four major sections: (1) a general section, collecting information about the person completing the questionnaire; (2) a task rating section where incumbents rated how often they performed each of 204 job tasks and where supervisors rated the importance of each of those tasks; (3) an equipment rating section in which each of 44 pieces of equipment were rated with respect to their frequency of use; and (4) a worker characteristics section in which each of 122 knowledges, skills, abilities, and personal characteristics were rated in terms of their usefulness for successful job performance.

The final job analysis questionnaires were administered to a random sample of 631 experienced incumbents (those on the job at least six months) and most of the supervisors in the participating agencies (251 supervisors). In addition, 65 incumbents and 83 supervisors completed the questionnaire a second time in order to determine its test-retest reliability.

Analysis of the completed questionnaires resulted in a statewide definition of the juvenile counselor "core job", as well as specific definitions of how the job is performed in each participating agency.

The core job was defined, based upon the task ratings from all participating agencies, as consisting of all job tasks listed in the job analysis questionnaire which:

1. were performed by half or more of the incumbents across most agencies,

OR

2. were rated as having medium or higher importance by the supervisors in most agencies.

Both of these indices were computed using a procedure which weighted all departments equally. This was done to ensure that the results were applicable to all participating departments and so that the results from large departments did not completely dominate. the study. Job tasks not meeting one of the above criteria were excluded from the core job definition - as were those tasks with medium or higher importance but which were rarely performed in most departments.

Apart from defining the core job for the juvenile counselor position, the job analysis data showed that the job analysis questionnaire was highly reliable over time, and that the juvenile counselor job did not vary from county to county. This latter finding was very important as it permitted the development of a uniform set of selection procedures to be applied to agencies throughout the State of California.

For a more detailed description of the research methodology involved in the job analysis, see the Job Analysis Technical Report, December, 1983 (available from the Board of Corrections).

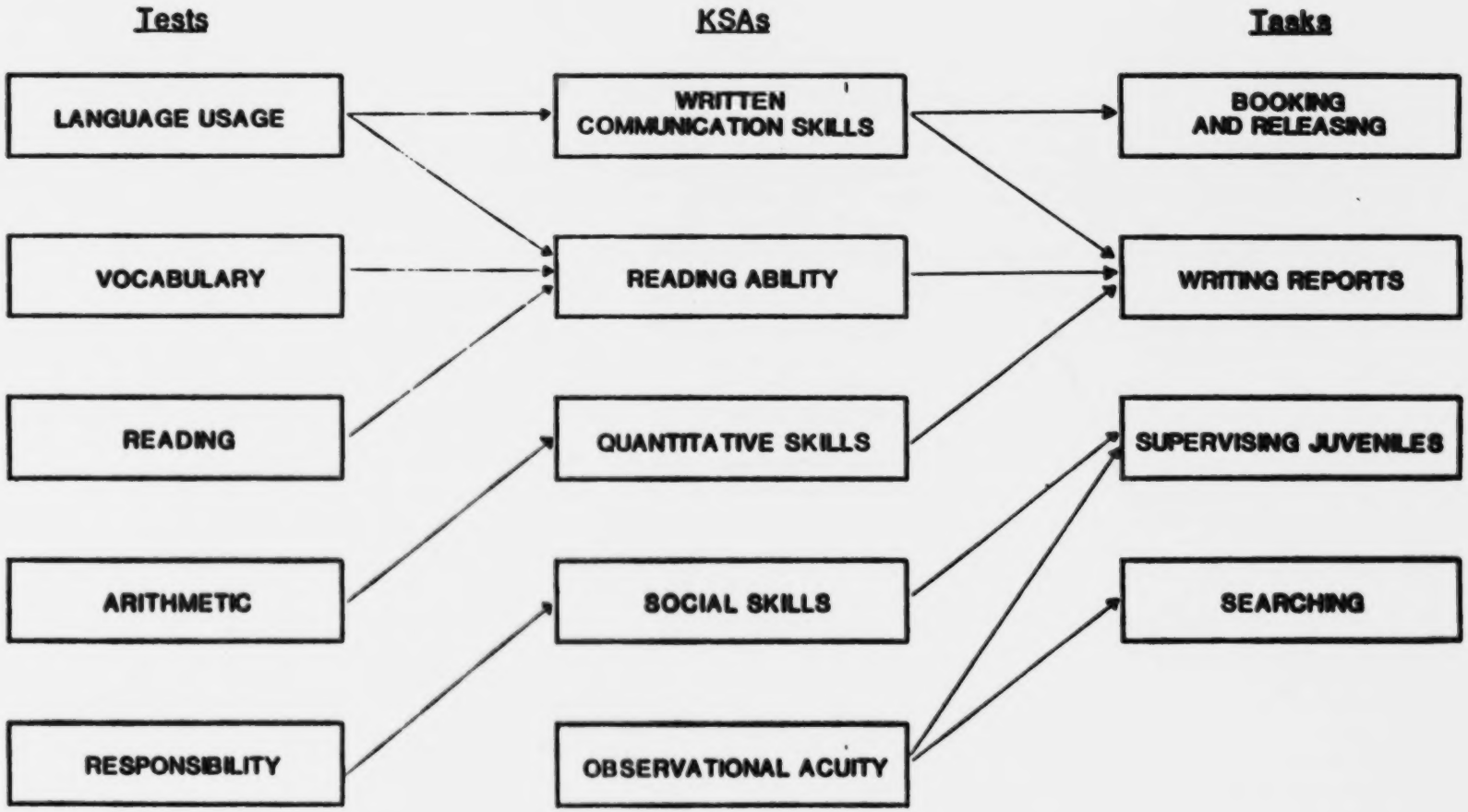
CHAPTER THREE: PREDICTOR DEVELOPMENT

Overview and Rationale

The end goal of predictor development was to obtain a set of tests (experimental selection battery) for administration to a sample of juvenile counselor incumbents (validation sample). These tests would comprise the building blocks from which the Juvenile Counselor Examination would be constructed. Because the experimental selection examination would form the basis of the final implemented examination, it was very important that the tests be highly indicative of candidates' ability to perform all important job tasks. It was necessary that the tests predict performance on the tasks that make up the job. However, tests do not measure task performance directly. What tests do measure are knowledges (e.g., knowledge of a facility's rules and regulations), skills (e.g., skill in adding or subtracting). abilities (e.g., ability to remain alert and aware), or personal characteristics (e.g., integrity). In order to link the tests to the job tasks, it was necessary to use the knowledges, skills, and abilities (KSAs) as intermediaries. Tests measure KSAs, not job tasks. Once it could be shown which KSAs were essential to the performance of which tasks, tests could be selected that measure those KSAs and, therefore, that predict task performance. Figure 3.1 shows the hypothetical relationships among a set of tests, KSAs, and tasks. The arrows show which tests are useful measures of which KSA and which KSAs are required for the successful performance of which tasks. Successful report writing, for example, requires three KSAs (Written Communications Skills, Reading Ability, and Quantitative Skills). Four tests are useful measures of one or more of these three KSAs. Because those tests are useful measures of the set of KSAs necessary for successful report writing, they can be used to predict report writing performance. In this project, the determination of which KSAs are important for specific tasks was made by juvenile counselor supervisors who rated the importance of every KSA for the successful performance of every job task. The usefulness of each of 72 tests as a measure of every KSA was rated by PDI staff psychologists experienced in the use of tests in making selection decisions.

Chapter 3 describes the methods used to link tasks and tests to KSAs. Also in Chapter 3 is a description of how several new tests were drafted, pretested, and refined for inclusion in the experimental selection battery.

Figure 3.1 Hypothetical Example of the Relationships Among Tests, KSAs, and Job Tasks



Linking KSAs to Job Tasks

As was noted one aspect of predictor development involved the linking of KSAs (directly measurable) to job tasks (ultimate criterion). Linking was accomplished by having job experts (juvenile counselor supervisors) rate the importance of each knowledge, skill, or ability for successful performance of the job tasks. Because there were 204 job tasks and 122 KSA items in the job analysis questionnaire, a reduction in the number of elements to be rated was necessary if the task was to be administratively feasible.

Reducing the number of KSAs to be rated was accomplished by combining them into broader groupings. First, PDI staff members rationally clustered the 122 individual KSAs into groups representing common characteristics (see Table 3.1). For example, several KSA items concerning aspects of facility-specific knowledge were grouped together. In this way, the 122 KSAs were combined into 26 worker characteristic clusters. This achieved a four-fold reduction in the number of KSAs to be rated, and at the same time, preserved the variety of skills and abilities important for successful job performance. The resulting worker characteristic clusters were reviewed by other PDI staff members and then "confirmed" via corroborative factor analysis.

Confirmation consisted of two separate oblique factorings of the worker characteristic data: one using incumbents' ratings of the usefulness of each KSA for successful job performance, the other using similar ratings from supervisors. The corroborative factor analysis began with the assignment of every KSA to a target factor (i.e., one of the 26 worker characteristic clusters). Then, based upon the intercorrelations of the KSAs, the appropriate matrix algebra equations were solved yielding a new matrix that showed the relationships among the KSAs and factors (Overall & Klett, 1971). This matrix was produced and examined to corroborate the rational clustering of the KSA. The clustering was corroborated when the KSAs correlated more highly with their target factor than they did with any of the other factors. Generally the worker characteristics were corroborated, but the results suggested that a few KSAs were more strongly related to other clusters. So long as the relocations made sense, KSAs were relocated and their correlations recomputed to corroborate the new structure.

As noted, corroboration was done separately using the JAQ data of incumbents and supervisors. It was found that the KSA items migrating, for example, from worker characteristic "A" to worker characteristic "F", when incumbents' data were analyzed, had a high probability of making the same migration when the supervisors' data were analyzed. This indicated that the rational clustering of KSAs into worker characteristics and subsequent minor revisions provided reliable results.

Table 3.1

Definitions of 26 Juvenile Counselor Worker Characteristics

A. Department-Specific Knowledge Knowledge of a facility's general rules and procedures, physical layout, shift schedules, security and communication systems; work and treatment programs; admitting, processing and releasing procedures; juvenile transport and discharge procedures; administrative and report-writing forms and procedures; and of available community resources.

B. Knowledge of California Juvenile Justice Knowledge of the history and philosophy of the juvenile court/justice system and juvenile probation process, of juvenile court forms, records and proceedings, juvenile rights, the California Welfare and Institutions Code and Youth Authority standards for institutions, and the legal rights and civil liability of facility staff.

C. Knowledge of Psychology Knowledge of crisis intervention, interviewing, diagnostic, and individual, group and family counseling principles and techniques; knowledge of group dynamics, human development, and personality theories; and skill in assessing juveniles' capabilities and security risk factors.

D. Street Knowledge Knowledge of narcotics and other drugs, gang behavior, street terminology, and various ethnic and social backgrounds.

E. Safety/First Aid Knowledge of emergency procedures, sources of emergency assistance, and standard institutional health and sanitation procedures; knowledge of, and skill in applying, first aid and CPR.

F. Operating Communications Equipment Skill in: use of radio, intercom, switchboard, and other oral communications equipment.

G. Oral Communication Skill in: giving clear verbal instructions/directions, testifying, and other oral communications.

H. Reading Skill in reading sufficiently to: understand facility forms, reports and manuals; understand/interpret laws, statutory codes, and court and other legal documents.

I. Written Communication Knowledge of report-writing procedures, grammar and spelling, and skill in completing forms and writing understandable and legible reports.

J. Quantitative Skill Skill in adding and subtracting.

Table 3.1 (continued)

K. Skill in Observing Behavior Skill in accurately observing, recalling and documenting behavior, visual details (such as identification marks). and other events; skill in noticing subtle changes in behavior patterns, and in staying alert to and understanding signs of potential problems or disturbance.

L. Driving Skill in driving automobiles, vans, trucks, and buses.

M. Observing and Searching Knowledge of observation and surveillance procedures, juvenile count and control procedures, and search procedures; skill in conducting pat- and strip-searches and in searching cells/dorms, vehicles, and other areas for contraband and weapons.

N. Self-Defense/Physical Restraint of Others Knowledge of, and skill in: application of self-defense, physical restraint, and riot control techniques and procedures; ability to apply the appropriate "use of force".

O. Organizing/Supervising Group Activities Knowledge of methods and techniques for organizing and leading group activities, of the rules of team and individual games/sports, and of specific crafts; ability to supervise group sports.

P. Social Skills Ability to maintain discipline without causing unnecessary tension; effectiveness in working with others and interacting with diverse groups; consistency and fairness in dealing with others; insight, sensitivity, and concern for others' feelings; tolerance for differences in values and beliefs; sense of humor.

Q. Motivation/Dependability Dependability and trustworthiness (e.g., coming to work on time, not abusing sick leave, assuming a fair share of the workload); willingness to put in extra time and effort; persistence in seeking information; integrity in adhering to professional ethics; commitment to a facility's goals and objectives.

R. Stress Tolerance Tolerance for the stress of working in a potentially dangerous environment; ability to maintain self-control when confronted with verbal harassment or threats; ability to stay calm and respond with composure in emergencies.

S. Dominance/Self-Esteem Assertiveness, self-assurance, and decisiveness; ability to influence and confront others, to enforce rules, to interact confidently with high-status individuals (e.g., judges), to make decisions where there is no standard procedure, and to delegate responsibilities and duties properly.

Table 3.1 (continued)

T. Non-Autonomy Ability to work under close supervision; tolerance of highly structured/regimented work environment; willingness to obey authority.

U. Flexibility Resourcefulness in thinking of alternative approaches (e..g., alternative treatment techniques, referral sources, etc.); ability to adjust to new and unfamiliar situations/duties; resilience in coping with setbacks.

V. Tolerance of Negative Aspects of the Job Tolerance of physically unpleasant work environments; ability to perform repetitive tasks effectively and remain alert and aware during long periods of inactivity.

W. Physical Abilities Upper body and leg strength; whole body coordination; manual dexterity and eye/hand coordination; endurance/stamina; and agility/flexibility

X. Vision, Hearing Daytime, nighttime, peripheral, and color vision; depth perception; and hearing acuity to be able to detect subtle changes in noise levels and quiet sounds such as whispering or scuffling.

Y. Investigative Ability Knowledge of investigation procedures, information sources, and laws pertaining to evidence; skill in gathering and evaluating information.

Z. Personal Organization Efficiency; ability to make good use of one's time, to meet deadlines and work effectively under time pressure.

Grouping Task Items into Task Categories

In the job analysis questionnaire, the 204 task items had been grouped into 15 broad categories of job responsibility (e.g., Searching and Securing). For the same reason that the KSA items were grouped into a fewer number of worker characteristics, it was necessary to group the task items into a fewer number of task categories. This reduction was accomplished by rewriting the individual task items in each general JAQ category into one to five intermediate categories, each consisting of task items related to similar behaviors or responsibilities (e.g., Obtaining details of charges from police report, see Table 3.2). The resulting 50 task categories (a four-fold reduction in number) were reviewed for completeness and were then used to link job tasks to the worker characteristics.

Task Category to Worker Characteristic Ratings

Fifteen juvenile counselor supervisors from four agencies rated how important the 26 worker characteristics were for the successful performance of every task category (approximately 1300 ratings per supervisor). On average, the raters were responsible for supervising 15 juvenile counselors, and had themselves been supervisors for an average of 7.5 years. They also had served as juvenile counselors for an average of 6.5 years. Ratings were made using a 5-point scale, indicating the importance of each worker characteristic for every task category (1 = not important, 2 = low, 3 = medium, 4 = high, and 5 = very high importance). Table 3.3 contains the average importance rating of each worker characteristic for every task category.

The reliability of these ratings was estimated using the intraclass correlation approach described by Shrout and Fleiss (1979). Twenty-six reliability estimates were computed, one for each worker characteristic rated. The ratings were reliable to the extent that the within-task category ratings variance was significantly less than the between-task categories variance. The specific model used to estimate the reliability of the ratings treated raters as a fixed effect (i.e., did not seek to generalize the results beyond those particular raters) and used the mean rating as the unit of analysis. The analysis sought only to estimate the reliability of the ratings as they were used in the study. Because the average rating would be used to link worker characteristics to task categories, the reliability of the average rating, rather than of the raters per se, was estimated.

The reliability of the ratings for each worker characteristic are shown in Table 3.4. Reliabilities range from .92 (Reading, Self-defense, and Physical Ability) to .43 (Motivation) with an average of .82, indicating that the ratings were made very reliably and therefore it is unlikely that the ratings would have differed appreciably should the supervisors have made the ratings a second time.

Table 3.2

50 Task Categories for Juvenile Counselors

Booking, Receiving, and Releasing

1. Obtaining details of charges from police report, reviewing booking forms and court documents.
2. Processing juveniles (by assessing medical needs, taking custody of belongings, observing behavior, assigning housing).
3. Verifying identity, performing necessary paperwork, and returning property before release.
4. Completing booking sheets, preparing forms, cards, file jackets.
5. Notifying parents of juvenile's status.
6. Observing/supervising juveniles while awaiting move to assigned housing.

Admissions/Orientation

7. Issuing clothing and bedding.
8. Briefing juvenile about facility, rules and procedures, discussing and having juvenile sign behavior/admission contract.
9. Answering juvenile's questions about facility, staff, juvenile justice system.

Escorting, Transporting

10. Transporting juveniles and equipment to desired location, with appropriate searching, surveillance, and restraints.

Supervising Juveniles -- daily care and custody

11. Supervising, inspecting, and instructing in room cleanliness and personal hygiene procedures.
12. Identifying and dealing with medical and psychiatric needs.
13. Supervising meal-taking, special diets, helping with food and equipment preparation and inspection.
14. Breaking up disturbances, carrying out appropriate disciplinary action, evaluating juvenile behavior for special programs.
15. Interacting with juveniles, interviewing, informing, responding to questions.
16. Conducting bedding exchange, sorting, laundering dirty linen or clothing.

Table 3.2 (continued)

Supervising Juveniles -- recreational or leisure time activities.

17. Planning, scheduling, and supervising sports, game and library activities, checking equipment in and out.
18. Coaching or participating in sports or game activities with juveniles.

Supervising Juveniles -- work details.

19. Inspecting and issuing equipment, instructing juveniles in safety procedures.
20. Supervising work assignments and clean-up .

Casework

21. Writing initial evaluation and treatment plan.
22. Counseling juveniles individually and in groups.
23. Evaluating progress and, with information from others, making further recommendations.

Record Keeping, Report Writing

24. Keeping and updating daily records of juvenile special status, movement, funds, and behavior.
25. Logging keys in and out, maintaining inventory, requisitioning repairs.
26. Preparing reports of incidents, disciplinary action, individual evaluations, and escapes.

Searching and Securing

27. Counting juveniles, verifying location, screening mail.
28. Routinely checking fences, equipment, patrolling hallways, perimeter.
29. Keeping inventory of keys and dangerous tools.
30. Investigating suspicious activities, searching people and areas, as necessary.
31. Searching juveniles, areas, for contraband.
32. Maintaining surveillance of juveniles.

Supervising Non-inmate Movement, Visitors

33. Admitting, screening, searching visitors and vehicles.
34. Supervising contacts.

Table 3.2 (continued)

Emergencies

35. Alerting all staff of emergency, evacuating juveniles, administering medical help.
36. Assisting other units, dispatching help.
37. Extinguishing, helping extinguish fires, conducting fire drills.

Communicating

38. Exchanging information with other shifts, central control, other personnel.
39. Making announcements, testifying in court, notifying parents of changes.
40. Providing information for visitors, regulatory agencies, parents/friends of juveniles.

Physically Demanding Tasks

41. Watching/listening for unusual activities.
42. Running, climbing, jumping, balancing oneself in pursuit of an escapee.
43. Physically subduing/separating violent or fighting juveniles.
44. Lifting, pushing, carrying, dragging heavy objects.
45. Walking or standing for long periods of time.

Miscellaneous Tasks

46. Dealing with accidents and/or crimes within facility (investigating, arresting, charging).
47. Keeping up with the job by attending meetings and training sessions, reading memos, log entries, and rules.
48. Carrying out court orders, assisting and instructing juveniles in schoolwork, voluntary, and vocational activities.
49. Testifying in court.

Supervising Other Detention Facility Personnel

50. Training, scheduling, and evaluating personnel, giving assignments.

Table 3.3 Average Importance Ratings of 26 Worker Characteristics for 50 Task Categories

Rating Scale:

1 = NOT IMPORTANT
2 = LOW IMPORTANCE
3 = MEDIUM IMPORTANCE
4 = HIGH IMPORTANCE
5 = VERY HIGH IMPORTANCE

JOB TASK CATEGORIES	Task No.	Department-Specific Knowledge	Knowledge of California Juvenile Justice	Knowledge of Psychology	Street Knowledge	Safety, First Aid	Operating Communications Equipment	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Observing and Searching	Self-Defense, Physical Restraint of Others	Organizing, Supervising Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Assertiveness/Self-Esteem	Non-Autonomy	Flexibility	Tolerance of Negative Aspects of the Job	Physical Abilities	Vision, Hearing	Investigative Ability	Personal Organization
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
I. Booking/Receiving/Releasing	1	4.2	3.9	2.1	2.1	1.9	1.7	3.0	4.7	3.0	2.1	2.0	1.0	1.8	1.7	1.3	2.1	3.1	2.5	1.9	2.1	2.3	2.4	1.6	2.5	3.4	2.3
	2	4.1	3.7	3.7	2.9	2.9	1.9	3.9	3.2	3.2	2.7	4.3	1.1	4.3	2.5	1.5	3.7	3.5	2.7	3.0	2.4	2.8	2.9	2.1	3.3	3.8	3.6
	3	3.8	3.0	1.8	1.7	1.7	1.7	2.7	3.7	3.3	2.3	2.1	1.0	2.1	1.2	1.1	2.6	3.1	2.1	2.3	2.1	2.1	2.2	1.5	2.7	3.7	3.1
	4	4.3	2.8	1.5	1.6	1.2	1.9	2.6	4.1	4.0	1.9	1.7	1.1	1.6	1.2	1.1	1.7	3.3	2.0	1.9	2.3	2.0	2.4	1.3	2.7	2.0	2.3
	5	4.1	3.5	2.6	1.7	1.4	3.2	4.1	3.0	2.1	1.2	1.5	1.0	1.1	1.1	1.0	2.9	3.0	2.9	2.3	2.2	2.7	2.7	1.2	2.4	2.0	2.6
	6	3.7	2.3	3.5	3.3	2.8	1.9	3.9	2.1	2.1	2.0	4.5	1.3	3.7	2.6	2.1	3.5	3.5	3.1	3.3	2.7	3.3	2.9	2.7	4.0	2.3	2.6
II. Admissions/Orientation	7	3.0	1.9	1.9	1.5	2.1	1.1	3.0	1.9	1.9	2.3	2.6	1.0	2.1	1.5	1.1	2.5	2.5	2.0	2.1	2.0	1.9	2.1	1.8	2.6	1.6	2.7
	8	4.3	3.2	3.3	2.9	2.1	1.5	4.5	3.8	2.7	1.9	3.5	1.2	2.4	1.7	1.6	3.6	3.3	3.1	2.9	2.5	3.2	2.5	1.7	3.1	2.1	2.9
	9	4.6	4.4	3.3	2.7	1.8	1.4	4.3	2.9	2.5	1.9	3.0	1.1	1.9	1.6	1.6	3.5	3.2	2.8	2.8	2.3	3.1	2.3	1.6	2.9	2.2	2.7
III. Escorting/Transporting	10	4.1	2.9	2.9	3.0	3.3	2.2	3.3	1.9	1.7	2.2	4.1	4.3	4.0	3.4	2.0	3.5	3.3	3.5	3.3	2.5	2.9	2.8	3.2	3.8	1.8	3.1
IV. Supervising Juveniles -- Daily care and Custody	11	3.3	2.8	2.6	2.2	2.9	1.4	4.2	1.9	2.0	1.7	3.5	1.0	3.9	1.7	2.3	3.4	3.4	2.5	2.7	2.5	2.8	2.5	2.0	2.8	1.9	3.1
	12	3.7	3.1	3.9	2.8	4.1	1.5	3.7	2.5	2.9	1.5	3.7	1.1	2.2	1.7	1.6	3.3	3.5	3.1	3.0	2.6	3.2	2.5	2.3	3.1	2.7	2.9
	13	3.4	2.6	2.3	2.0	2.9	1.3	3.2	2.1	2.0	2.2	2.8	1.2	2.1	1.4	2.2	3.3	2.9	2.1	2.5	2.7	2.3	2.3	1.9	2.8	1.9	2.7
	14	4.2	3.8	3.9	3.5	3.4	2.1	4.2	2.2	3.3	1.8	4.4	1.1	3.3	4.3	2.3	3.8	3.5	4.1	3.9	2.9	4.3	3.9	4.1	3.4	3.8	3.3
	15	4.0	3.6	3.8	3.4	2.3	1.4	4.5	2.4	2.7	1.8	3.9	1.1	2.5	1.9	2.5	4.1	3.3	3.4	3.3	2.3	3.3	2.9	2.1	2.9	2.8	3.3
	16	2.7	2.0	1.8	1.6	2.3	1.1	2.5	1.6	1.5	2.6	2.2	1.0	2.6	1.3	1.7	2.3	3.0	1.9	1.8	2.1	1.9	3.3	1.9	2.1	1.9	2.4
V. Supervising Juveniles -- Recreational or Leisure Time Activities	17	3.3	2.2	2.9	2.5	3.1	1.5	3.5	2.5	2.5	2.8	3.1	1.1	2.9	2.0	4.4	3.5	3.4	2.6	2.9	2.6	3.6	2.6	2.8	2.8	2.1	3.7
	18	3.1	2.0	2.8	2.6	3.5	1.4	4.1	1.9	1.9	2.1	3.8	1.2	2.5	1.9	4.3	3.9	3.4	2.5	3.3	2.5	3.7	2.4	3.5	3.1	1.7	3.0
VI. Supervising Juveniles -- Work Details	19	3.5	2.1	2.5	2.0	3.9	1.7	3.7	2.0	2.3	2.1	3.3	1.4	2.5	1.7	2.7	3.3	3.5	2.5	2.5	2.7	2.5	2.7	2.3	2.7	2.1	2.7
	20	2.8	2.0	2.7	2.3	3.4	1.3	3.6	1.9	1.9	1.8	3.9	1.1	2.9	1.7	3.7	3.9	3.3	2.7	3.4	2.2	3.3	2.9	2.7	3.2	1.7	3.3
VII. Casework	21	3.3	2.5	3.6	2.5	1.4	1.1	1.7	3.5	4.4	1.6	3.6	1.0	1.6	1.3	1.3	2.3	3.0	2.0	2.0	2.1	3.1	2.4	1.2	2.1	2.3	3.2
	22	3.4	2.6	4.2	3.9	1.9	1.3	4.4	2.4	2.3	1.5	3.9	1.1	1.9	1.7	3.5	4.3	3.5	3.5	3.6	2.2	4.0	2.5	1.8	2.8	2.5	3.1
	23	3.1	2.4	3.5	2.9	1.6	1.2	2.8	3.1	3.7	1.8	3.7	1.1	2.2	1.3	1.5	3.1	3.3	2.3	2.9	2.6	3.3	2.6	1.5	2.4	3.1	3.3
VIII. Record Keeping, Report Writing	24	3.3	2.1	1.9	1.7	1.4	1.5	2.2	3.9	4.2	2.8	2.5	1.0	1.5	1.1	1.1	1.7	3.1	2.1	1.7	2.7	2.3	2.7	1.5	2.1	2.3	2.9
	25	3.3	1.9	1.1	1.1	2.1	1.5	1.9	2.6	3.2	2.2	1.5	1.0	1.6	1.3	1.2	1.9	3.0	2.1	2.1	2.0	2.8	2.5	1.5	2.1	1.7	2.8
	26	3.5	2.9	2.7	2.5	1.7	1.7	2.9	3.7	4.5	2.1	4.1	1.2	2.4	1.4	1.7	2.5	3.5	2.3	2.7	2.6	3.1	2.4	1.8	2.3	3.9	3.6
IX. Searching and Securing	27	3.3	2.3	1.7	2.8	1.9	1.5	2.6	3.1	2.3	3.3	3.0	1.1	2.7	1.5	1.7	2.3	3.1	2.6	2.3	2.8	1.9	3.0	1.8	3.3	2.2	2.8
	28	3.3	1.7	1.3	1.3	2.4	1.3	2.1	1.5	1.5	1.6	2.9	1.0	3.7	1.8	1.3	1.8	3.1	2.5	2.1	2.8	2.1	3.3	2.5	4.0	2.9	2.8
	29	3.0	1.5	1.4	1.3	2.7	1.1	1.9	1.9	2.3	3.0	2.2	1.0	2.5	1.3	1.1	1.4	2.9	2.1	2.3	2.8	1.7	2.5	1.5	3.0	2.2	2.7
	30	3.4	2.3	2.8	3.2	2.5	1.6	3.3	1.8	2.0	1.9	4.1	1.3	4.1	2.5	1.6	3.2	3.6	3.3	3.3	2.6	3.6	3.1	2.9	3.9	4.3	3.1
	31	3.5	2.3	2.9	2.9	2.3	1.3	2.6	1.7	1.6	1.8	3.4	1.0	4.3	1.9	1.7	3.1	3.2	2.5	3.1	2.9	2.5	2.9	2.7	3.5	3.9	3.1
	32	3.5	2.3	2.8	3.3	2.5	1.7	3.1	1.6	1.5	1.8	3.7	1.0	3.7	1.9	1.7	2.8	3.2	3.0	2.6	2.7	3.1	3.2	2.4	3.7	2.9	2.9

BEST COPY AVAILABLE

Table 3.3 (continued)

Rating Scale:

1 = NOT IMPORTANT
2 = LOW IMPORTANCE
3 = MEDIUM IMPORTANT
4 = HIGH IMPORTANCE
5 = VERY HIGH IMPORTANCE

JOB TASK CATEGORIES	Task No.	Department-Specific Knowledge	Knowledge of California Juvenile Justice	Knowledge of Psychology	Street Knowledge	Safety, First Aid	Operating Communications Equipment	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Observing and Searching	Self-Defense, Physical Restraint of Others	Organizing, Supervising Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Flexibility	Tolerance of Negative Aspects of the Job	Physical Abilities	Vision, Hearing	Investigative Ability	Personal Organization
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
II.																											
Supervising Inmate Movement, Visitors	33	3.6	2.6	2.2	2.7	2.2	1.9	3.1	1.8	1.8	1.7	3.6	1.1	4.0	1.8	1.6	3.5	3.1	2.9	3.3	2.7	3.3	3.1	2.0	3.5	3.1	3.0
	34	3.2	2.7	3.1	3.3	1.8	1.7	3.3	2.4	2.7	1.9	3.7	1.3	2.6	1.7	1.5	3.7	3.7	3.1	3.4	2.8	3.3	2.7	1.9	3.3	2.8	3.2
III.																											
Emergencies	35	4.5	2.7	2.7	1.9	4.9	4.2	4.7	2.3	2.5	2.5	3.4	1.1	2.3	2.5	2.1	3.4	3.0	4.2	3.7	3.1	3.6	3.1	3.4	3.8	2.5	3.7
	36	4.3	2.1	2.2	1.6	3.5	2.9	4.5	2.0	2.3	2.1	3.0	1.1	2.7	3.5	1.8	3.0	2.7	3.8	3.3	3.1	3.5	3.2	3.2	3.6	2.8	3.6
	37	4.5	2.3	1.9	1.5	4.0	2.6	3.5	1.8	1.7	2.0	2.9	1.0	1.7	1.9	2.7	2.7	3.3	3.8	3.3	3.0	3.5	3.2	3.7	3.5	2.6	3.5
III.																											
Communicating	38	3.7	2.3	2.1	2.3	2.3	2.4	4.3	3.1	3.3	2.1	3.1	1.0	2.1	1.3	1.5	3.5	3.8	3.0	3.3	2.8	2.9	2.8	2.1	2.5	2.4	3.2
	39	3.7	2.7	2.5	2.0	1.5	2.0	4.7	2.5	2.5	1.6	2.9	1.0	1.6	1.2	1.3	3.7	3.5	2.9	3.3	2.3	3.0	2.4	1.5	2.8	2.5	3.1
	40	4.4	4.1	2.5	2.1	2.3	1.7	4.5	2.9	2.9	1.9	2.9	1.0	1.8	1.3	1.3	3.9	3.2	2.7	3.3	2.3	3.1	2.7	2.0	2.4	2.1	2.9
XIII.																											
Physically Demanding Tasks	41	3.1	2.0	2.3	2.7	2.6	2.4	2.1	1.5	1.7	1.6	4.2	1.3	3.7	2.3	1.7	2.5	3.3	2.3	2.8	2.4	2.7	3.4	1.9	4.5	3.7	2.7
	42	2.8	2.3	1.9	2.3	2.8	1.7	2.1	1.3	1.3	1.3	2.5	1.2	2.6	3.9	1.7	1.9	3.3	3.3	2.6	2.4	3.1	3.3	4.3	4.0	2.5	2.6
	43	3.9	3.5	3.3	3.4	3.9	2.1	2.9	1.3	1.4	1.5	3.8	1.2	3.0	4.5	2.3	2.8	3.5	4.2	3.9	3.1	3.5	3.7	4.5	3.8	2.3	3.0
	44	2.4	1.9	1.3	1.3	3.7	1.1	1.7	1.2	1.5	1.6	1.5	1.2	1.4	1.7	1.5	1.7	2.5	2.5	1.9	2.1	1.9	3.4	4.2	1.9	1.1	1.7
	45	1.8	1.5	1.4	1.2	2.6	1.2	1.6	1.2	1.6	1.6	1.7	1.1	1.5	1.2	1.4	1.6	2.6	2.4	1.9	2.3	2.0	3.2	3.8	1.9	1.6	1.6
XIV.																											
Miscellaneous Tasks	46	4.3	4.4	2.8	2.6	3.0	1.7	3.2	2.9	3.3	1.9	4.1	1.1	3.3	2.6	1.9	3.1	3.4	3.3	3.3	2.7	3.3	3.1	2.0	3.5	4.5	3.5
	47	2.2	2.7	2.2	1.7	1.9	1.7	3.6	4.0	3.8	1.9	2.6	1.7	1.3	1.3	1.3	2.9	3.5	2.8	2.8	2.6	3.3	2.7	1.5	2.9	2.1	3.9
	48	3.7	3.0	2.5	2.0	2.1	1.3	3.2	3.9	2.6	2.6	3.1	1.2	1.2	1.1	2.4	3.3	3.2	2.1	2.2	2.6	2.5	2.0	1.5	2.5	1.5	2.4
	49	3.7	4.4	2.9	2.6	1.3	1.1	3.9	2.6	2.2	1.4	3.5	1.0	2.0	1.1	1.3	3.4	3.4	3.3	3.4	2.5	3.4	2.9	1.4	2.9	2.7	2.6
XV.																											
Supervising Other Detention Facility Personnel	50	4.2	1.5	3.7	2.2	3.1	2.9	4.2	4.1	4.2	3.4	3.9	1.1	2.6	2.1	2.7	3.2	3.8	3.4	3.2	3.6	4.1	2.9	2.6	3.5	3.2	4.4

Setting Linkage Criteria

Once the task category to worker characteristic ratings were collected and determined to have been reliable, it was necessary to document the results and to define the level of importance (i.e., average rating) at which a worker characteristic would be considered linked to a task category. For purposes of summarizing the strength of the task category to worker characteristic linkages, the 5-point rating scale was trichotomized into low (average rating of 1.0 - 1.5), moderate (1.5 - 3.5), and high (3.5 - 5.0). Table 3.4 summarizes the number of task categories linked to each worker characteristic at each of the three linkage strengths. As can be seen, for approximately 2/3 of the worker characteristics the task category linkages were of a single strength. For the remaining 1/3, the linkages were approximately evenly divided between two strengths.

It was desired to set the criterion for linking a worker characteristic to a task category as high as possible, so that the tests developed to measure each linked worker characteristic would provide accurate predictions of task performance. Developing a test to measure a worker characteristic that is only weakly linked to the task category the test is to predict, would not result in a selection battery likely to provide accurate predictions of job performance. Initially, the criterion was set at 4.0; that is, only those worker characteristics with average importance ratings of 4.0 or higher would be considered linked to a task category. As will be discussed however, it was necessary to relax this criterion to ensure that all of the tasks were linked to at least one predictor. At the time these data were being collected and analyzed, the core job had not been defined (see Job Analysis Report for details). As a result, the relative importance of the 204 job tasks was not known. Rather than risk overlooking a task category that might later have been found to be an important, or even a critical aspect of the juvenile counselor job, the linking criterion was adjusted so that every task category, therefore every task, would be linked to at least one worker. characteristic for which a test could be developed.

Linking Tasks to Worker Characteristics

In much the same way that worker characteristics were linked to task categories, it was necessary to link the tests being considered for inclusion in the experimental selection battery to the same set of worker characteristics (recall that it is through the worker characteristics that tests were linked to the tasks they were to predict). The first step in this process was to compile a list of tests having potential to predict the job performance of juvenile counselors. Based upon research reports located during the literature review, the job analysis materials and results, and the experience of PDI staff members, a list of 72 tests was compiled.

For each test, a one-page description was constructed describing the characteristic it measured, its reliability and, if available, a summary of its validity for similar jobs.

Table 3.4
Reliabilities and Ratings Summary
for Juvenile Counselors

Worker Characteristic	Rating rel.	Strength of Task Relationships[b] Low	Mod.	High
A. Department-specific knowledge	.83	0	26	24
B. Knowledge of California juvenile justice	.90	1	39	10
C. Knowledge of psychology[a]	.90	6	38	6
D. Street knowledge	.88	7	41	2
E. Safety/first aid	.89	5	36	9
F. Operating communications equipment	.87	22	26	2
G. Oral communication[a]	.90	1	28	21
H. Reading[a]	.92	4	36	10
I. Written communication[a]	.91	5	38	7
J. Quantitative skill[a]	.76	7	43	0
K. Skill in observing behavior[a]	.88	1	27	22
L. Driving[a]	.91	48	1	1
M. Observing/searching	.89	6	34	10
N. Self-defense/restraint of others	.92	18	29	3
O. Org/supervising group activities[a]	.90	18	28	4
P. Social skills[a]	.86	2	34	14
Q. Motivation/dependability[a]	.43	0	41	9
R. Stress tolerance[a]	.81	0	45	5
S. Dominance/self-esteem[a]	.83	0	45	5
T. Nonautonomy[a]	.45	0	50	0
U. Flexibility[a]	.82	0	40	10
V. Tolerance of negative job aspects[a]	.60	0	48	2
W. Physical ability[a]	.92	7	37	6
X. Vision and hearing[a]	.82	0	38	12
Y. Investigative ability	.87	2	42	6
Z. Personal organization[a]	.71	1	42	7

[a]Before-hire worker characteristic
[b]See text (p. 19) for full descriptions

Eleven psychologists from PDI and its sister organization Personnel Decisions Research Institute rated the degree to which each of the 72 tests would be a useful measure of each of the juvenile counselor worker characteristics. The full set of 26 worker characteristics was narrowed to 18 by deleting those characteristics which, on the basis of the job analysis data, were found as not being necessary before hire. This was done because any worker characteristic not necessary at the time of hire (i.e., that could be trained on the job) would not be an appropriate selection standard and, therefore, linking tests to those worker characteristics would have served no useful purpose. In order that no worker characteristic be prematurely rejected from consideration, this narrowing of the list of worker characteristics was done as conservatively as possible. Before-hire worker characteristics were defined as those having at least one constituent KSA which, according to at least 40% of Juvenile counselor supervisors, should be developed before hire (i.e., at least 40% of supervisors rated the KSA 0 or 1 on the 0 to 4 "when developed" JAQ scale).

Ratings of the usefulness of each test as a measure of each worker characteristic were made on an 8-point scale ranging from "not at all useful" (rating of 0), through "almost useless" (ratings of 1 or 2), "of some use" (ratings of 3, 4, or 5), "highly useful" (ratings of 6 or 7), to the worker characteristic being "entirely measured" by the test (rating of 8). Table 3.5 shows the average usefulness of each test as a measure of each of the 18 before-hire worker characteristics.

The reliability of the test to worker characteristic ratings was determined using the same methods described for determining the reliability of the task category to worker characteristic ratings. The reliability analyses indicated that, with the exception of the Physical Ability worker characteristic, the ratings were all quite reliable (range = .72 to .97, average = .86). The reliability of the Physical Ability ratings (intraclass r = .08) at first glance appears quite dismal, but this apparent unreliability is a statistical artifact rather than an indication of unreliable ratings. The intraclass correlation estimate of reliability is an index of the relative amount of variation in the ratings data. As the data in Table 3.5 reveal, there was almost no variation in the psychologists' ratings of the usefulness of the 72 tests as measures of the physical ability worker characteristic (i.e., no tests were rated as useful measures of physical abilities). Because there was essentially no variation in the ratings, the intraclass correlation must be near zero. Nevertheless, the ratings were made reliably in the sense that all raters agreed as to the usefulness (lack of usefulness, actually) of the 72 tests as measures of juvenile counselor physical abilities.

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 18 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Organizing, Supervising	Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Personal Organization	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
1. EAS Ver. Comp.	3.3	5.0	5.0	5.4	1.4	1.3	.1	2.1	1.4	.4	.4	1.2	.1	1.0	.4	1.4	.1	0	
2. EAS Num. Ab.	1.5	.4	1.3	1.2	7.5	.3	.5	.9	.1	.2	.1	.4	.2	.4	.2	.5	.1	0	
3. EAS Visual S & A	.3	.1	1.4	1.0	1.2	2.3	1.3	.5	.1	.2	.1	.2	.4	1.4	.9	.1	1.4	0	
4. EAS Space Vis.	.1	.2	.4	.3	1.1	1.0	2.4	.9	.1	.1	.1	.1	.2	.6	.2	.1	1.4	.2	
5. EAS Ver. Reas.	3.4	3.3	3.8	3.6	.9	1.1	.3	1.2	1.1	.1	.2	.6	.4	.6	.1	1.0	.1	.1	
6. EAS Symb. Reas.	2.1	1.1	1.5	1.8	2.0	.6	.3	.4	.3	.1	.1	.4	.3	.4	.1	1.1	.1	.1	
7. DAT Abst. Reas.	2.7	1.4	2.0	1.6	2.3	1.2	.4	1.0	.4	.1	.1	.1	.3	.4	.1	1.1	.4	.1	
8. DAT Lang. Usage	2.1	4.2	4.1	6.3	1.0	.3	.1	.8	.7	.1	0	.4	.2	.4	.1	.3	.1	.1	
9. DAT Ver. Reas.	3.8	3.6	3.6	3.8	.9	1.1	.1	.5	.7	.1	.1	.4	.3	.4	.1	.9	.1	.1	
10. EEI Follow D.	.8	1.8	1.7	.8	1.0	.7	1.5	1.3	.2	.4	.3	.2	2.2	1.1	.9	.3	.9	.1	
11. EEI Tables & Gra.	.8	.5	1.4	.6	1.3	.9	.4	.2	.1	.1	.1	.2	.9	.6	.4	.2	1.3	.1	
12. ETS Hid. Fig.	.2	.1	.4	.4	.3	2.3	1.3	.3	.1	.1	.1	.2	.2	.3	.2	1.6	1.6	.1	
13. ETS Card Rotat.	.2	.1	.5	.3	.6	1.9	1.3	.7	.1	.1	.1	.1	.1	.3	.1	1.1	1.1	.1	
14. ETS Map Mem.	.4	.2	.6	.3	.5	3.1	1.7	.5	.1	.1	.1	.1	.2	.4	.1	.2	1.2	.1	

[a] Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 18 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Organizing, Supervising	Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Personal Organization	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
15. ETS Bldg. Mem.	.3	.1	.4	.2	.4	3.0	1.4	.4	.1	.1	.1	.1	.1	.4	.1	.2	1.2	.1	
16. ETS Follow Dir.	.8	2.0	2.2	.9	.7	.7	.8	1.2	.1	.1	.1	.1	1.6	.6	.8	.6	.3	.1	
17. Watson-Glaser	3.4	2.9	4.1	3.2	1.4	1.5	.4	.8	1.2	.3	.1	.6	.4	1.3	.1	.9	.1	.1	
18. CO Reading	2.0	2.3	7.0	2.8	1.0	.6	.4	1.2	.3	.1	.1	.1	.3	.5	.1	.3	.2	.1	
19. Otis-Lennon	3.2	2.8	3.4	3.0	3.4	1.5	.4	1.4	1.0	.3	.1	.4	.3	1.1	.1	1.0	.2	.1	
20. Wesman Verb.	3.2	3.0	3.9	3.4	1.8	1.4	.3	1.5	1.1	.3	.1	.2	.3	1.2	.1	.9	.1	.1	
21. CAB Mn	1.3	.9	1.4	1.3	.8	2.0	1.0	.8	.3	.1	.1	.1	.6	.2	.4	.2	.4	.1	
22. CPI Do	.8	2.3	.3	.9	.4	1.1	.4	4.6	3.2	1.2	1.6	6.6	1.6	1.4	1.0	1.4	.1	.1	
23. CPI Sy	1.3	2.4	.5	.3	.2	1.4	.1	4.1	5.3	.8	.8	2.6	1.1	.4	.5	1.4	.1	.1	
24. CPI Wb	1.2	1.0	.3	.1	.1	.7	.1	2.2	3.4	2.8	2.9	3.5	1.9	.8	3.0	1.4	.1	.1	
25. CPI Re	.9	.4	.3	.4	.1	.7	.7	1.6	1.7	5.9	1.9	1.4	4.4	2.0	2.7	1.0	.1	.1	
26. CPI So	.5	.4	.3	.2	.1	1.4	.9	1.9	2.4	5.3	2.0	1.4	4.0	1.2	3.2	1.5	.1	.1	
27. CPI Sc	1.3	1.0	.3	.2	.1	.5	1.1	1.8	2.4	3.2	4.3	1.6	3.0	1.8	2.3	1.6	.1	.1	
28. CPI To	1.7	1.5	.4	.3	.3	1.4	.5	2.7	4.4	1.7	2.2	1.7	3.3	.6	2.2	2.8	.1	.1	

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 18 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Organizing, Supervising Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Personal Organization	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
29. CPI Ac	1.8	1.4	1.3	1.3	.8	.9	.4	1.8	1.4	3.4	1.4	1.0	3.9	2.9	2.4	2.3	.1	.1
30. CPI Fx	1.0	1.2	.4	.4	.2	.4	.5	.9	1.3	1.0	1.4	.7	2.8	1.1	2.0	4.4	.2	.1
31. CPI Ie	3.0	2.4	2.4	2.3	1.9	1.6	.6	1.7	1.4	1.7	1.3	1.3	1.4	1.9	.7	1.7	.1	.1
32. CPI Ai	1.8	1.7	1.5	1.6	1.3	.9	.4	1.2	1.4	1.7	.8	1.2	1.7	1.2	.8	2.0	.1	.1
33. MMPI L	.8	.6	.2	.1	.1	.6	.1	.7	1.4	.8	1.4	1.6	1.2	.2	.7	.7	.1	.1
34. MMPI F	.6	.8	.8	.3	.2	.7	.1	.9	1.5	1.6	2.5	.8	.6	.5	.7	.1	.1	.1
35. MMPI K	.2	.5	.2	.1	.1	.8	.1	.7	1.9	1.6	1.6	1.1	1.0	.1	.8	.9	.1	.4
36. MMPI 1	.8	.4	.2	.1	.1	1.4	.3	.8	2.2	1.7	2.4	1.4	.2	.5	1.0	.3	.2	.1
37. MMPI 2	.4	.7	.2	.1	.1	.7	.5	1.4	3.0	1.5	2.6	2.3	0	.6	1.1	.9	0	.1
38. MMPI 3	.4	0	.1	0	.2	.4	.3	1.0	1.9	1.8	3.5	.9	.6	.2	2.0	.4	0	0
39. MMPI 4	.3	.8	.1	0	0	.4	1.2	1.4	2.7	3.3	1.3	1.6	2.8	.4	1.5	.3	0	0
40. MMPI 5	.6	.3	.3	.2	.2	1.3	.2	.5	1.5	.8	.5	1.4	.7	.2	.4	.7	0	0
41. MMPI 6	.6	.2	.1	0	0	.8	.2	.7	1.1	.3	.9	.9	.1	0	.7	0	0	0
42. MMPI 7	.8	1.3	.2	.7	.1	1.5	0	1.9	1.7	2.1	2.7	2.8	2.0	3.9	2.0	1.8	0	0

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 18 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Organizing, Supervising Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Personal Organization	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
43. MMPI 8	.6	.9	.1	0	0	.9	.2	1.4	2.4	1.5	2.0	1.8	1.7	.8	1.9	.2	0	0
44. MMPI 9	.5	.7	.4	.3	.3	.7	.8	2.9	2.0	2.1	2.2	2.2	.6	1.1	.7	.7	0	0
45. MMPI 0	.2	1.9	.1	.1	.1	1.1	0	3.9	4.6	1.4	1.7	3.0	.2	.2	.2	1.1	0	0
46. MMPI 0-H	.7	.2	0	0	0	.2	.7	1.6	2.1	.9	2.8	.9	2.3	1.0	1.3	1.2	0	0
47. GPI Coop.	.4	.7	.1	.1	.1	.8	.2	2.1	4.4	1.5	2.2	1.4	1.2	0	1.1	.9	0	0
48. GPI Conscient	.1	0	0	.4	.2	.6	1.4	.1	.4	2.3	1.4	1.2	4.4	.5	1.0	1.4	0	0
49. GZTS Gen. Act	.4	.8	.2	.3	.2	0	.4	2.5	1.6	1.5	.7	1.9	.9	.9	1.0	.3	0	.2
50. GZTS Ascend.	.4	2.4	.2	.2	.2	.4	0	4.5	3.5	1.4	1.4	5.4	1.6	.4	.5	.2	0	0
51. GZTS Emot. Stab	.6	1.4	.1	.1	.1	.5	.4	1.7	2.4	2.1	3.9	1.6	.7	.8	2.3	1.4	0	0
52. GZTS Objectiv.	.6	.9	.2	.2	.2	.8	.2	2.0	2.3	1.3	4.1	2.4	1.1	.2	2.2	1.6	0	0
53. GZTS Pers. Rel.	.4	.4	.1	.1	.1	.4	.2	1.6	3.6	1.5	1.2	1.4	1.6	.3	1.5	1.2	0	0
54. GZTS Restraint	.4	.5	.1	.2	.2	.7	1.4	1.1	1.1	3.5	2.2	1.0	1.6	2.7	1.3	.6	0	0
55. DPQ Unfrien	.1	.1	0	0	0	.8	.4	1.3	3.2	1.4	2.0	1.0	1.4	.2	1.8	1.0	0	0
56. DPQ Authorit.	.4	.1	0	0	0	.6	.4	.9	1.4	2.9	1.1	.7	3.7	1.1	2.3	1.4	0	0

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 18 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Organizing, Supervising	Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Personal Organization	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
57. DPQ Stress R.	.6	.8	.1	.1	.1	.4	.8	1.4	1.4	1.8	5.3	1.9	.7	.5	2.8	1.6	0	0	
58. DPQ Aggression	.2	.2	0	0	.5	.2	1.6	2.1	3.4	2.2	2.3	1.6	1.3	.4	1.3	1.4	0	0	
59. DPQ Impulsive	.5	.8	0	0	0	.2	.7	2.0	1.7	3.0	1.5	.5	2.1	4.0	1.6	2.2	0	0	
60. DPQ Control	.3	.1	0	0	.2	.2	.4	1.5	1.1	3.2	1.5	.4	2.8	4.7	1.5	1.6	0	0	
61. DPQ Tradition	.3	.1	0	0	0	.2	.4	.8	1.2	3.6	.4	1.3	3.2	1.2	3.0	2.2	0	0	
62. APQ	.1	1.2	0	0	0	.4	.2	2.3	2.9	.3	4.9	1.4	.4	.3	2.7	1.1	0	0	
63. BQ Soc. Inr	.4	1.8	0	0	0	.4	0	3.6	4.0	.2	1.2	2.8	.4	.1	.4	.5	0	0	
64. BQ Agg/Indep	.1	1.7	0	.2	0	.4	0	3.0	2.3	.3	.5	4.0	1.7	.2	1.1	1.0	0	0	
65. BQ Soc. Des.	.2	.4	0	.2	0	.4	.4	.8	1.6	1.0	1.3	2.3	1.7	.2	1.7	.3	0	0	
66. BQ Soc. Lead.	.1	1.8	.2	.1	.2	.3	0	5.1	4.4	.9	1.2	3.9	.8	.2	.5	.2	0	0	
67. BQ Adjust.	.5	.6	0	0	.2	.4	.3	1.4	2.5	1.2	2.5	2.9	.6	.2	1.2	.4	0	.1	
68. BQ Exp. Neg.	.4	.5	0	.1	0	.1	0	.6	2.4	.5	1.4	.8	1.2	0	1.2	.1	0	0	
69. Self-Esteem	.6	1.0	.3	.3	.3	.2	.2	1.5	1.9	.7	1.8	3.5	.2	.4	.9	.4	0	.2	
70. Empathy	1.3	.8	0	0	0	2.4	.1	2.3	4.5	.5	.5	1.5	.8	.2	.9	1.9	0	0	

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 18 Before Hire Worker Characteristics

Test/Scale	Knowledge of Psychology	Oral Communication	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Driving	Organizing	Supervising Group Activities	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Non-Autonomy	Personal Organization	Tolerance of Negative Aspects of the Job	Flexibility	Vision, Hearing Abilities	Physical Abilities
71. Co Interest	.7	.5	0	0	0	1.3	0	2.3	2.3	1.8	2.1	1.1	1.9	.2	1.9	1.5	0	0	
72. Occup. Role	.8	.7	.1	0	.1	.1	0	1.8	1.7	1.2	.7	.7	2.2	.8	1.7	.6	0	.3	

Table 3.5 (continued)

California Psychological Inventory

22. Dominance
23. Sociability
24. Well-being
25. Responsibility
26. Socialization
27. Self control
28. Tolerance
29. Achievement via conformance
30. Flexibility
31. Intellectual efficiency
32. Achievement via independence

Minnesota Multiphasic Personality Inventory

33. Lie scale
34. Infrequency
35. Validity
36. Social introversion
37. Hypomania
38. Schizophrenia
39. Psychasthenia
40. Paranoia
41. Masculinity-femininity
42. Psychopathic deviate
43. Conversion hysteria
44. Depression
45. Scale 0
46. Scale 0-H

Gordon Personal Inventory

47. Cooperativeness
48. Conscientiousness

Guildford-Zimmerman Temperament Survey

49. General activity
50. Ascendancy
51. Emotional stability
52. Objectivity
53. Personal relations
54. Restraint

Table 3.5 (continued)

Multidimensional Personality Questionnaire

55. Unfriendly world
56. Authoritarianism
57. Stress Reaction
58. Aggression
59. Impulsiveness
60. Control
61. Traditionalism

62. Activity Preference Questionnaire

Owens Biographical Questionnaire (unpublished)

63. Social introversion
64. Aggressiveness/independence
65. Social desirability
66. Social leadership
67. Adjustment
68. Expressing negative emotion

69. PDRI Self Esteem

70. Hogan's Empathy

71. Correctional Officers' Interest Blank

72. PDRI Occupational Role Preferences.

To summarize the results of the test to worker characteristic ratings, the 8-point scale was reduced to four categories. Tests whose average usefulness rating fell between 0.0 and 0.5 were said to have no relationship to the worker characteristic, tests with ratings between 0.5 and 2.5 were said to have a low relationship, between 2.5 and 5.5 a moderate relationship, and tests with an average rating of 5.5 or above were said to be highly related to the worker characteristic. Table 3.6 presents the number of tests having each degree of relationship to each of the 18 before-hire worker characteristics. As can be seen, most worker characteristics were judged moderately related to several tests. There were two exceptions to this finding: Driving and Physical Abilities, both of which were essentially unrelated to any of the 72 tests.

Linking Tests to Task Categories

Once the task category to Worker Characteristic and worker characteristic to test ratings were made, the stage was set to link tests to task categories through the intervening worker characteristics. This linking of tests to task categories required the simultaneous examination of both sets of ratings. The strategy adopted for linking tests to tasks was to set criteria that were as stringent as possible, while at the same time ensuring that at least one test was linked to each task category. Procedurally, each task category was considered in turn and all of the worker characteristics whose average ratings met or exceeded the specified criterion (see Table 3.7) were linked to that task category. Each of the worker characteristics so linked were themselves examined and all of the tests whose average usefulness rating met or exceeded the second criterion were linked to the task. For example, for the first task category (Obtaining details of charges from police report), one worker characteristic (Reading) had an average importance rating of 3.5 or higher. For Reading two tests (EAS Verbal Comprehension and CO Reading) received average usefulness ratings of 5.0 or higher.

Initially it was hoped that two criteria could be set, one to link all task categories to worker characteristics, the other to link all tests to worker characteristics. As Table 3.7 indicates, it was occasionally necessary to decrease the magnitude of one or the other criteria in order that at least one test be linked to each task category. Once all test to task category linkages were made, the list of linked tests was revised to eliminate duplicates (due to worker characteristics being linked to more than one task), yielding the final set of 12 tests (see Table 3.8) that comprise the experimental selection battery for juvenile counselors.

Table 3.6

Reliabilities and Ratings Summary of Test to Worker Characteristic Ratings

Before-hire Worker Characteristic	Strength of Task Relationships[a] None	Low	Med.	High	Rating rel.
A. Knowledge of psychology	28	36	8	0	.88
B. Oral communication	22	43	7	0	.89
C. Reading	49	15	7	1	.97
D. Written communication	51	13	7	1	.96
E. Quantitative skill	49	21	1	1	.95
F. Skill in observing behavior	23	47	2	0	.72
G. Driving	47	25	0	0	.74
H. Organize/supervise	5	57	10	0	.87
I. Social skills	15	40	17	0	.91
J. Motivation/dependability	25	36	10	1	.90
K. Stress tolerance	22	38	12	0	.89
L. Dominance/self-esteem	19	42	10	1	.90
M. Non-autonomy	24	38	10	0	.86
N. Tolerance of negative job aspects	21	45	6	0	.78
O. Flexibility	22	48	2	0	.74
P. Vision and hearing	64	8	0	0	.82
Q. Physical abilities	72	0	0	0	.08
R. Personal Organization	34	33	5	0	.89

[a]See text for full descriptions.

Table 3.7

Linkage Criteria and Summary for Test to Worker Characteristic and Worker Characteristic to Task Category Ratings

Task Category[a]	Task Category to Worker Characteristic		Worker Characteristic To Test	
	Criterion	N Linked	Criterion	N Linked
1	3.5	1	5.0	2
2	3.5	6	5.0	4
3	3.5	1	5.0	2
4	3.5	2	5.0	4
5	3.5	1	5.0	1
6	3.5	6	5.0	4
7	2.7	2	5.0	1
8	3.5	4	5.0	4
9	3.5	2	5.0	2
10	3.5	5	5.0	2
11	3.5	2	5.0	1
12	3.5	4	5.0	3
13	3.0	2	5.0	2
14	3.5	10	5.0	7
15	3.5	4	5.0	2
16	3.0	2	5.0	2
17	3.5	5	5.0	3
18	3.5	6	5.0	3
19	3.5	2	5.0	3
20	3.5	4	5.0	3
21	3.5	4	5.0	4
22	3.5	9	5.0	8
23	3.5	3	5.0	2
24	3.5	2	5.0	4
25	3.2	2	5.0	3
26	3.5	5	5.0	6
27	3.0	6	5.0	5
28	3.0	3	5.0	2
29	3.0	2	5.0	1
30	3.5	4	5.0	2
31	3.2	3	5.0	2
32	3.5	2	3.0	2
33	3.5	3	5.0	1
34	3.5	3	5.0	3
35	3.5	7	5.0	6
36	3.5	6	5.0	4
37	3.5	6	5.0	2
38	3.5	3	5.0	4
39	3.5	3	5.0	4
40	3.5	2	5.0	2
41	3.5	2	3.0	2

Table 3.7 (continued)

Task Category[a]	Task Category to Worker Characteristic Criterion	N Linked	Worker Characteristic To Test Criterion	N Linked
42	3.5	2	5.0	0[b]
43	3.5	8	5.0	5
44	3.5	1	5.0	0[b]
45	3.5	1	5.0	0[b]
46	3.5	3	4.5	1
47	3.5	5	5.0	7
48	3.5	1	5.0	2
49	3.5	2	5.0	1
50	3.5	12	5.0	10

[a]See Table 3.2 for Task Category Labels.

[b]Task Categories 42, 44, and 45 were strongly linked to worker characteristics (Vision and Hearing, Physical Ability) that were not linked to any tests.

Predictor Development

As was detailed above, PDI and PDRI staff psychologists rated the usefulness of 72 tests of cognitive or noncognitive abilities as measures of the worker characteristics necessary to successfully perform the juvenile counselor job. Analysis of the ratings showed them to be very reliable and indicated that 12 of the 72 tests (see Table 3.8) were likely to be useful predictors of job performance. Six of the tests were measures of noncognitive abilities. The remaining six were measures of cognitive abilities. The contract between the Board of Corrections and PDI allowed for the use of existing noncognitive tests in the final selection battery, but required all cognitive ability tests to be newly developed so that the copyright to the tests could be assigned to the Board of Corrections. In accordance with the contract, the six noncognitive measures were permissible for use in the selection battery "as is", contingent only upon appropriate permissions and royalty agreements being negotiated with the test publishers. The remaining six tests, being measures of cognitive ability, could not be used in the final battery, and so required the development of new tests designed to measure the same abilities.

Development of the new cognitive ability tests involved three steps. First, items were drafted and preliminary test booklets assembled. Second, the newly constructed tests were pretested in a sample of college students (administered with the marker tests they were designed to replace). Third, data from the pretest were used to refine the tests into a form ready for use in the experimental selection battery. Each of the three steps is detailed below.

Test Construction. In order that the new tests parallel as closely as possible the format and content of the marker tests they were to replace, the first step in test construction was to outline a test development plan for each test. The development plans (see sample in Figure 3.2) described the ability to be measured, the number and type of items to be included, whether the test was intended to be a speed or power test, the proposed time limit for the test, and any other comments that might be useful in guiding construction of the test. Once the test development plans were completed, test construction proceeded similarly for each of the five tests. Only five tests were developed because two of the marker tests (ETS Map Memory, ETS Building Memory) were so similar in content that a single test was developed to replace both of these marker tests.

Table 3.8

Cognitive and Noncognitive Tests for the Experimental Selection Battery

Marker	Test Description
Cognitive Ability	
EAS Verbal Comprehension	A 30-item vocabulary test with a five-minute time limit designed to measure the ability to use words in oral and written communications.
EAS Numerical Ability	A 75-item arithmetic test in three parts with a total test 10-minute time limit. The test measures a person's facility with the arithmetic operations of addition, subtraction, multiplication, and division using problems combining integers, decimals, fractions, and percentages.
DAT Language Usage	A 50-item test in which persons must identify which of four sections of a sentence contains an error of punctuation, grammar, or capitalization. The time limit for this test is 25 minutes.
ETS Map Memory and ETS Building Memory	Each is a 24-item test with 12-minute time limits (6 minutes for each of two parts) designed to measure one's "ability to remember the configuration, location, and orientation of figural material". The tests differ only in the nature of the material to be memorized (map sections versus building locations).

Table 3.8 (continued)

Marker	Test Description
CO Reading	A 50-item reading test that measures the ability to read and remember what was read. The test consists of short passages each followed by several questions.
Noncognitive Tests	
CPI Dominance	A 46-item subscale of the California Psychological Inventory that identifies individuals who behave in a dominant, ascendant manner and who, in interpersonal situations, can take initiative and exercise leadership in a forceful, self-confident manner.
CPI Sociability	A 36-item subscale of the California Psychological Inventory that indicates a person's tendencies to be outgoing and participative and to seek out and enjoy social contact without being shallow or superficial.
CPI Responsibility	A 42-item subscale of the CPI measuring a person's tendency to believe that life is best governed by reason and to behave in a responsible, self-disciplined manner.
CPI Socialization	An item subscale of the CPI indicating a person's tendency to be role-respecting, prudent, circumspect, and responsive to what others feel and think.
MPQ Stress Response	A 26-item subscale of the Multi-dimensional Personality Questionnaire that indicates persons who are able to put fears and worries out of their minds easily; who quickly get over upsetting experiences, and who are not troubled by emotional turmoil or feelings of guilt.
MPQ Control	A 24-item scale of the MPQ measuring a person's tendency to be cautious, careful, rational and sensible, and indicating persons who like to anticipate events and plan their activities.

Figure 3.2

Test Development Plan for Numerical Ability Test

Ability Measured
Mathematical knowledge and computational ability.

This test is to measure knowledge of the four basic mathematical operations (addition, subtraction, multiplication, and division). Applicants will be required to apply this knowledge to simple problems involving whole numbers, fractions, decimals, and percentages.

Marker Test
Employee Aptitude Survey Test 1.

Test Format
Test should consist of 50 multiple choice items, each with four distractors. This is intended to be a speed test with a time limit, excluding instructions, of five minutes.

Comments
This test is intended to roughly parallel the EAS marker but with fewer items and a slightly different item mix. Distractors should be devised to be the results that would obtain from common errors. For example, a problem such as 1/3 x 1/2 might have as distractors 1/5, 2/5, 2/6, and the correct answer, 1/6. In this way, the test will tap both the knowledge (what to do) and computational (actually doing it) aspects of numerical ability.

Using the test construction plans as blueprints, PDI staff members drafted 2 to 2-1/2 times the number of items needed for the final tests. The surplus of items ensured that the tests could be significantly revised and still contain enough items to be reliable and valid.

Once items were drafted, two preliminary forms of each test were assembled and scoring keys developed and checked. Item distractors were arranged so that the correct response occupied each possible position (i.e., A, B, C, D, or E) with approximately equal frequency. Each test was then reviewed by PDI staff members for obviously biased or unfair items and the scoring keys rechecked for accuracy. The resulting preliminary tests were printed in sufficient quantity for pretesting. Preliminary test construction yielded two forms of each of the five cognitive ability tests to be used in the experimental selection battery.

Pretesting the Preliminary Tests. Arrangements were made with instructors at four-year and community colleges to permit their students to serve as participants in the pretest sessions. In return the students received extra credit toward their class grade. Each preliminary test was administered, under the same conditions to prevail during actual test use, to groups of 15-30 students. No student completed more than two of the preliminary tests. Table 3.9 presents the number of students completing each of the preliminary tests. Two of the tests listed in Table 3.9 were not specifically targeted for the juvenile counselor position. The tests were developed as potential predictors of one of the two other jobs studied in this project. They are included in Table 3.9 because, although not targeted to the juvenile counselor job, they were part of the experimental selection battery administered to juvenile counselors. A single version of the experimental selection battery was prepared for use with all three jobs. This was considered necessary because validation testing for the three jobs occurred simultaneously. It was critical to prevent any accidental cross administration of, say, the probation officer tests to juvenile counselors.

During pretesting, students were administered the appropriate marker test (under the time limits and instructions provided by the publisher), followed by both forms of the preliminary tests. For the preliminary test, students were instructed to indicate the item they were working on at the end of the proposed time limit, but were then permitted to continue working until they had completed all items. This procedure allowed the time limits to be gauged as well as permitting the collection of data about every test item. Because the data were used to compute the item statistics that were in turn used to select the final items, it was important to ensure that psychometrically

Table 3.9

Pretest Results and Item Statistics for the Experimental Selection Battery

Test	Number of Students Tested	Avg. Item Diff.	Item-Test Corr.	N Items	Rel.[b]	Validity[c]
Observational Memory	30	.90	.10	30	.44	.31
Numerical Ability	19	.74	.57	35	.92	.75
Vocabulary	21	.66	.37	35	.83	.82
Language Usage	29	.78	.51	35	.90	.88
Reading Comprehension	30	d	d	70	d	d
Analogies[a]	16	.78	.62	35	.93	.69
Verbal Reasoning[a]	25	.57	.26	35	.80	.72

[a]Not targeted for the juvenile counselor job -- see text (pp. 39) for details.

[b]Internal consistency via KR-20.

[c]Correlation with marker test.

[d]Data unavailable due to method of revision -- see text (pp. 43) for details.

sound items were not eliminated from the final tests simply because they appeared near the end of a preliminary test. Such elimination might have occurred had the pretesting time limits been so short as to prevent the students from attempting all of the items.

Item Analysis and Revision. Using the data collected in the pretest sessions, three statistics were computed for every preliminary test item. The difficulty of each item (proportion of persons passing the item) was computed as was the correlation of each item with the total score. Item-total correlations were computed as the point-biserial correlation between an item (scored dichotomously as pass or fail) and the score received on the sum of the remaining items. Finally, the frequency with which each distractor was chosen was determined.

The item statistics were used to select items to appear in the experimental selection battery. Prior to this step, the average number of items completed in the proposed time limit was computed for each test. These results showed that the students were working the items quickly enough to permit a switch from speed to power measures of each of the cognitive abilities. Speed measures are those where the constituent items are so easy that, given sufficient time, everyone should be able to get every item correct. For speed tests, time limits are set that are much shorter than is necessary for most people to complete all test items. Speed tests are predicated upon the idea that people with more or higher levels of the ability being measured will be able to work the items faster. Thus, a higher score on a speed test is an indication of better ability. On the other hand, power tests have time limits that permit most persons to complete all items. On power tests the items vary in difficulty, with some being so easy that nearly everyone can get them correct while others are much more difficult.

To summarize, speed tests depend upon time limits to distinguish among persons of varying ability while power tests depend upon item difficulty. The advantage of power tests, especially in a project such as this, lies in their independence from time limits.

In order for any test to give an accurate measure of ability, or in this case the potential to be a successful juvenile counselor, the test must be administered under as nearly identical conditions as possible every time it is used. With speed tests, especially those with relatively short time limits, it is very difficult to ensure that the time limits are rigidly followed during every administration. Because they have no strict time limits, power tests are more desirable for projects such as this. As a result, each test in the experimental selection battery (with two exceptions discussed below) contained approximately the average number of items completed in the pretesting

time limit. The total time allowed for the resulting battery was then doubled. Thus, the final time limits were set so that virtually every person would be able to complete all items within the allotted testing time. In this manner, the tests were transformed from highly speeded to power tests, with an accompanying reduction in administrative complexity.

Once the desired number of items was determined for each test, the pretest data were used to screen items and to reject those that were not of appropriate psychometric quality. Several factors were considered in the course of this screening. For example, items for which the keyed alternative correlated negatively with total scores, or where the item-total correlation for the keyed alternative was less than that for a distractor, were rejected. No initial screening was done on the basis of item difficulty. The items remaining after screening formed the final pool of psychometrically sound items from which item selection was made.

In order to ensure an overall difficulty level appropriate for the intended job applicant population, items were selected based upon their item difficulties. For example, five items with difficulties between 1.0 and .90, five with difficulties between .90 and .80, and so on, were chosen for each test. Within difficulty levels, the items with the highest item-total correlations were selected..

Once the desired number of items had been selected, the reliability (KR-20) and validity (correlation with the published marker test) of each test were determined. If these were satisfactory, test construction was complete. If not, the correlation of each item with the marker test total score was computed (for all items in the final item pool) and this index used as the basis for item selection. That is, within a specific range of item difficulty, the items which correlated most highly with the marker total score were selected. Table 3.9 summarizes the psychometric properties of the tests that make up the experimental selection battery.

As was true for the preliminary tests, items in the final tests were arranged so that no patterns of correct answers were present and so that the correct alternative appeared in each possible position with approximately equal frequency. Finally, items were arranged (approximately) in ascending order of difficulty.

Concern that the tests might prove too difficult for the applicant population led to the inclusion of five additional items in each test. All of the additional items had item difficulties of 1.0 in the pretest sample. These items were included to ensure that each test had sufficient breadth to prevent restricting the range of test scores due to floor effects. "Floor effects" refers to a situation in which the items of a test are all too difficult for the test taking population. If this occurs, scores on the test will all be very low and there will

be a large number of persons who receive the same score. A large number of tied scores reduces the variance of the scores and puts a limit on the maximum possible validity of a test. In order not to limit the potential validity of the tests, several very easy items were added to each test.

Special Revision Methods. One of two tests not revised as described above was the Observational Memory Test. Rather than selecting a subset of the 30 items in the preliminary version of this test, it was decided to retain the test intact. Data on the pretest time limits showed that the test was easily completed in a minimal amount of time, therefore little was to be gained by shortening the test. The time limit results, coupled with item statistics indicating that the items were too easy, led to a reduction in the suggested time limits. Rather than having two minutes to study the stimulus figures to be memorized, only one minute was allowed. The amount of time allowed to complete the ten items in each test section was also reduced from three to two minutes. In addition to the changed time limits, the items were made more difficult by increasing the complexity of the stimuli to be memorized and by requiring more subtle discriminations between correct and incorrect alternatives. No other revision of the test was undertaken and it remained the only test to depend upon speed rather than power to distinguish among candidates.

The other test for which a variation in the revision procedures was made was the Reading Comprehension test. Pretest data for the reading marker test (CO Reading) revealed that the test was too easy for the pretest sample, leading to very little variation in test scores. Much the same was true for two of the three parts of the preliminary reading test. As a result, the item statistics were judged too unreliable to serve as a guide to test revision. One section of the preliminary test (Part 2) in particular was clearly not working as intended. This part of the test required subjects to make inferences and draw conclusions from minimal amounts of information. The problem with this part of the test was that no clearly justifiable key could be agreed upon and subjects were about equally divided in their choice of distractors.

The experimental selection battery Reading Comprehension test was made up of the six reading passages that PDI staff judged to be most appropriate for the juvenile counselors. The 10 to 12 items following each passage were an approximately 75%-25% mix of "search and find" and reasoning items. The former required the recall of answers to factual questions, while the latter required simple inferences and conclusions to be drawn from the material presented in the passages. These reasoning items are the only remaining vestige of Part 2 of the preliminary scale. Enough items were included in each passage to allow the test to be further refined (based on item statistics and using the procedures described above) using data collected during the validation phase of the project.

FINAL TEST BATTERY

The newly developed cognitive ability tests were assembled in three booklets. The first booklet contained the Observational Memory test. This test required strict adherence to time limits and so was kept separate from the other tests. The second booklet contained the Reading Comprehension test. A time limit of 30 minutes was set for this booklet. On the basis of the pretest data, 30 minutes was judged to be sufficient to allow virtually all incumbents to complete all test items. The test consisted of six passages, each with 10 to 12 items. Other than the 30-minute limit for the entire booklet, incumbents were not constrained in the amount of time they could spend on any one passage.

The third test booklet contained the Numerical Ability, Language Usage, Vocabulary, Analogies, and Verbal Reasoning tests. The Analogies and Verbal Reasoning tests, while not targeted specifically to the juvenile counselor job, were administered as part of the experimental selection battery. This was done for administrative convenience -- permitting administration of the same test battery to the validation samples for all three jobs covered in the project. Because testing was to occur simultaneously for all jobs, it was feared that separate test batteries might be misadministered. To prevent such an occurrence, a single battery containing the union of tests developed for all three jobs was assembled and administered to juvenile counselors as well as to probation officers and corrections officers.

Each test in Booklet 3 contained 35 items (175 items in all). A time limit of 60 minutes was set for the entire booklet. This allowed ten minutes per test (twice the estimated time necessary to complete 35 items) and left an additional ten minutes of testing time. Incumbents were free to divide the 60 minutes among the five tests as they saw fit. However, there were prominent warnings that no more than ten minutes should be spent working items for any given test until all tests had been attempted.

Noncognitive Tests

Noncognitive marker scales likely to predict job performance were identified at the same time and using the same methods as described for the cognitive marker tests. Six of the seven tests thus identified are shown in Table 3.8. The publisher of each test was contacted to secure permission to use their test in this project and to ascertain whether the Board of Corrections would be granted permission to score the tests. For one of the marker tests (Guilford-Zimmerman Temperament Survey) it seemed unlikely that the necessary permission would be obtained. The Guilford-Zimmerman Temperament Survey was omitted from the experimental selection battery.

Although not identified as a potentially useful scale in the analysis conducted to identify noncognitive marker tests, the Correctional Officers' Interest Blank (COIB) was selected for inclusion in the experimental selection battery based on its proven performance in similar studies. An additional noncognitive test (Hogan's Empathy) targeted for the other positions was also included in the experimental selection battery.

With the exception of 18 items from the COIB, all of the items from the noncognitive tests were in true/false format. These items were edited to change all occurrences of the pronoun "I" to "you". These changes necessitated other minor editorial changes so that the items read well. A few items were omitted because of their invasive nature. These items were potentially self-incriminating (e.g., "You have never been in trouble with the law") or dealt with topics too personal for use in an employment setting (e.g.. "You have been in trouble one or more times because of your sex behavior"). Editing was kept to a minimum. STC staff members also reviewed the items for biased or objectionable items, leading to the deletion of five items, as well as several minor editorial changes in item content.

All noncognitive items were assembled into one booklet with two sections. The first section contained the 228 true/false items from the noncognitive tests. Items from the different scales were mixed together unsystematically. The second section contained the remaining COIB items. These items required incumbents to pick from among three alternatives the statement they would like most and the one they would like least. Based on an estimated response rate of five items per minute (12 seconds per item) a time limit of one hour was set for the noncognitive items booklet.

Sample Items Booklet

To provide the incumbents in the validation sample an opportunity to become familiar with the cognitive ability items in the experimental selection battery, a booklet with illustrative items from each cognitive ability test was assembled. For each test the purpose of the examination was briefly explained, two or three sample items given with the correct answers explained, and eight to ten practice items (with answers) given. This booklet was distributed to all incumbents tested and was the prototype of a similar booklet to be made available to future job applicants as a preview of the Juvenile Counselor Examination.

CHAPTER FOUR: CRITERION DEVELOPMENT

A criterion-related validation strategy was used in this study. In a criterion-related approach, tests are valid if test scores can be demonstrated to be statistically related to a reliable measure of job performance. The agencies participating in this study used different methods of assessing the performance of their incumbents. Thus, it was not possible to use existing performance ratings as the criterion for judging the validity of the experimental tests. Even if a standard method of assessing incumbent performance had been available, it would not have proved useful because professional standards mandate that criterion-related studies include a relevant and reliable criterion measure. Existing job performance data were of unknown relevance and unknown reliability. Thus, it was necessary to develop new measures of job performance to be used as the criterion for assessing the validity of the predictors.

To ensure the relevance of the criterion measures, job analysis data were used to guide criterion development. These data provided a clear picture of the most important aspects of the juvenile counselor job and thus provided an ideal starting point for criterion development. The goal of criterion development was to create a series of scales that fully cover the relevant aspects of the juvenile counselor job. These scales would then be used by supervisors to rate the job performance of the incumbents who completed the experimental selection battery.

A critical incidents approach was taken to criterion development. In this approach, juvenile counselor supervisors wrote behavioral incidents representing examples of poor, average, and excellent performance for critical aspects of the juvenile counselor job. These incidents, when analyzed and summarized, provided the benchmarks that were used by other supervisors to rate the performance of incumbents in these important areas. Details of the critical incident methodology are given below.

Collecting Critical Incidents

Supervisors were recruited to participate in this phase of the study in such a way that large and small counties from both northern and southern California were represented. Participating supervisors were intimately familiar with the duties of juvenile counselor incumbents and were articulate and able to express themselves well in writing. A total of 39 supervisors attended 1/2-day workshops conducted by a PDI psychologist. During the workshops, the critical incident process was explained and the critical incidents written.

To guide their thinking about the job, supervisors were given a summary of the job analysis consisting of descriptions of the task categories developed for use in linking worker characteristics to task categories (See Chapter Three).

The task categories were explained to the supervisors who were then provided forms (See Figure 4.1) on which to record their critical incidents. These forms gave general instructions about writing critical incidents and provided space for the supervisors to: 1) describe the circumstances leading up to each incident; 2) describe what the incumbent did during the incident; 3) rate the incumbent's performance on a 1 (extremely ineffective) to 9 (extremely effective) point scale; and 4) indicate into which (if any) of 10 general task categories the incident fit.

Each supervisor was able to generate approximately 11 incidents yielding a total of 442 critical incidents from the 39 supervisors. These incidents provided the raw material from which development of the criterion measures proceeded.

Revising and Retranslating the Critical Incidents

The incidents were edited to emphasize the behaviors involved, to separate double incidents, and to ensure that all incidents were described in a similar format. PDI staff members rationally grouped the edited incidents into categories based upon the behaviors involved in each incident. The goal was to create categories which included only those incidents revealing varying levels of effective performance for similar or related job tasks. Eleven such categories of incidents were created and are described in Table 4.1. To corroborate the categories and to ensure that each incident was correctly assessed as to its degree of effectiveness, the 442 incidents were "retranslated" by juvenile counselor supervisors. During retranslation a second set of supervisors was given a list of the edited incidents and descriptions of the 11 categories. The supervisors were asked to place each incident in the most appropriate category. Retranslation was necessary to be sure that from the perspective of persons familiar with the job, each incident was correctly grouped and to ascertain the degree of effectiveness each incident represented. For reasons of administrative feasibility, the 442 incidents were assembled into two booklets, each with approximately 250 incidents (the booklets contained overlapping incidents to permit the data collected for each booklet to be equated). The booklets contained detailed instructions for the retranslation task and provided descriptions of the 11 performance categories to be used in making the ratings. Each booklet contained incidents believed to belong in each of the 11 categories.

Figure 4.1
SAMPLE PERFORMANCE INCIDENT FORM

Thinking back, focus your attention on any specific thing that you saw a person do that made you think of him/her as a highly effective worker, an average worker, a very ineffective worker, or any shade of effectiveness in between. In other words, think of a specific example of job behavior that gave you some feeling for how well the individual was doing his/her job. (Please do <u>not</u> record the names of the people involved.)

1. **What were the circumstances leading up to the incident?**

2. **What did the individual do that made you think he/she was a good, average, or poor performer?**

3. **How effective do you feel this example of performance was (Circle a number.)**

1	2	3	4	5	6	7	8	9
extremely ineffective		ineffective		about average		effective		extremely effective

4. **What, if any, general category would the above action fit into?**

Table 4.1

Juvenile Counselor Performance Dimensions
Used During Retranslation

A. MAINTAINING SECURITY

Maintaining appropriate close supervision over wards to prevent escapes, self-injuries, accidents, fights, other negative incidents; adhering to appropriate security procedures when escorting wards; safeguarding facility keys and potentially dangerous tools and supplies; verifying the identity of visitors and of wards receiving passes or releases.

B. SEARCHING

Exercising thoroughness and care in conducting required searches and inspections (of wards, visitors, mail, dorms, halls, facility yard area, etc.); initiating special searches when appropriate; taking appropriate action in follow-up to searches/inspections.

C. ALERTNESS TO POTENTIAL PROBLEMS

Being alert to signs of potential danger or disturbance; noticing sounds or subtle changes in wards' behavior or conversation that might indicate some type of trouble (escape plans, contraband possession/use, gang conflict, etc.),. and taking appropriate steps to prevent the trouble; recognizing signs of health problems, suicide risk, assaults, etc., and taking appropriate action to protect the well-being of wards; investigating suspicious activities, incidents, and situations.

D. KEEPING/CHECKING RECORDS

Performing all paperwork necessary for the correct and efficient operation of a facility: preparing all appropriate reports; reviewing court documents, case files, reports from other shifts, etc; interpreting court orders correctly; keeping accurate logs and inventories; keeping up-to-date on paperwork and reports; writing reports that are clear, complete, accurate, and concise.

Table 4.1 (continued)

E. HANDLING EMERGENCIES/CRISIS
Remaining calm and taking appropriate actions in emergency or crisis situations such as injuries, suicide attempts, fires, escapes, rioting, physical fights between wards or attacks upon staff, or other serious threats; sounding and responding to alarms with proper procedures; enlisting and providing appropriate assistance; using appropriate first aid, self-defense, or restraint procedures.
F. DEALING WITH HARASSMENT
Responding calmly and appropriately to verbal harassment, arguing, intimidation attempts, resistance, or other uncooperativeness from wards; responding with reasoning and patience, maintaining control, applying appropriate consequences, and taking appropriate disciplinary action versus responding fearfully or with foul language, threats of violence, or excessive use of force); handling confrontations in ways that reduce, rather than increase, conflict and tension.
G. DEALING WITH EMOTIONALLY DISTRESSED WARDS
Showing sensitivity and consideration in dealing with emotionally distressed, withdrawn, or self-destructive wards; using effective listening and counseling skills to identify and deal with underlying problems; making appropriate treatment and intervention recommendations.
H. DAY-TO-DAY INTERACTIONS WITH WARDS
Clearly and thoroughly communicating rules, policies, expectations, and consequences; enforcing rules and treating wards fairly and consistently; displaying concern for their health and general well-being; respecting their feelings, rights, and privileges; gaining their cooperation and respect; teaching, supporting, and rewarding positive behavior; listening and responding appropriately to their questions, concerns, complaints, and requests; providing appropriate assistance in working out problems.

Table 4.1 (continued)

I. INTERACTING WITH PARENTS

Notifying wards' parents/guardians of all pertinent developments; informing them of facility rules, procedures, and policies; establishing cooperative relations; respecting their feelings, rights, and privileges; providing counseling when appropriate to improve their effectiveness in dealing with the ward; dealing effectively with parents/guardians who are angry, upset, or resistant to security rules and procedures.

J. INTERACTING WITH OTHER STAFF

Getting along with co-workers, supervisors, and other law enforcement and criminal justice personnel; establishing and maintaining cooperative working relations; communicating effectively; keeping other staff completely informed of things they should know; following through on commitments to other staff; assisting, coaching, and supporting other staff as appropriate.

K. COMMITMENT TO THE JOB

Demonstrating commitment to the job and respect for a facility's policies and procedures: carrying out (versus disregarding) instructions from supervisors; putting in more time and effort than required versus just doing enough to get by on the job; doing things on one's own initiative to improve facility programs or organize special activities for wards versus not pulling one's fair share of the workload, for example, by coming in late or refusing to deal with certain wards; demonstrating professionalism in all work-related activities.

Forty-two juvenile counselor supervisors were selected to categorize the incidents into one of the 11 possible job performance categories and to rate the degree of effectiveness each incident represents. Effectiveness ratings were made on a 9-point scale. Twenty-four supervisors completed the first booklet and 18 completed the second. These supervisors represented 11 agencies and were responsible, on average, for supervising 15.1 juvenile counselor incumbents. They had been supervisors for an average of 5.9 years and the 37 who had also served as juvenile counselor incumbents reported an average of 6.9 years experience.

Developing Behavioral Summary Scales

For every retranslated incident, the proportion of supervisors who placed the incident in each category was computed, as was the mean effectiveness rating and the standard deviation of the ratings about that mean. These data were used to revise the categorization and to select incidents to appear as behavioral examples of poor, adequate, and excellent performance. Based upon the retranslation data, the category "Dealing with Harassment" was dropped and its items reassigned to the "Handling Emergencies/Crises" category. This was necessary because too few incidents to permit construction of a separate rating scale were categorized in the "Dealing with Harassment" category. The retranslation data also indicated that the "Interacting with Parents" and "Interacting with Other Staff" categories should be combined. Finally, the "Keeping/Checking Records" category was renamed as "Record Keeping/Report Writing".

Incidents within each category were sorted into those indicating poor, adequate, and excellent performance and used to construct 9 rating scales. Each scale (See Figure 4.2) provided a description of the performance category to be rated along with a 7-point rating scale. Ratings of 1 or 2 were labeled "needs improvement", ratings of 3, 4, or 5 were labeled "fully adequate", and ratings of 6 or 7 were labeled "excellent". Three to five incidents were abstracted to provide examples of the kind of behaviors that should result in a given performance rating. Instructions to raters were developed to emphasize that the behavioral examples were not checklists, but rather provided to guide the raters thinking about typical kinds of behaviors and therefore the level of performance of each incumbent being rated.

Other Rating Scales

<u>Worker Characteristic Ratings</u>. In addition to the 9 behavioral summary scales previously described, the 26 worker characteristics created during predictor development (see chapter three) were transformed into 7-point rating scales. Ratings of 1 or 2 indicated "low" standing (among the lowest one-third of employees), 3, 4, or 5 indicated "average" (in the middle one-third), and ratings of 6 or 7 indicated a "high" standing (upper one-third) on that worker characteristic.

Figure 4.2
Sample Behavioral Summary Scale

A. MAINTAINING SECURITY

Adhering to appropriate security procedures when escorting and supervising wards; maintaining sufficiently close supervision over wards to prevent escapes, self-injuries, accidents, fights, other negative incidents; safe-guarding facility keys and potentially dangerous tools and supplies; verifying the identity of visitors and of wards receiving passes or releases.

1 2	3 4 5	6 7
Needs Improvement	Fully Adequate	Excellent
Tends to be negligent about maintaining strict security. For example: - Leaves wards unsupervised. - Gets so involved in conversation or recreational activities with a few wards that s/he becomes inattentive to other wards' activities. - Allows escape opportunities when transporting wards. - Fails to conduct required head counts, to check on missing wards, or to verify identity of visitors.	Follows established security procedures. For example: - Makes sure that wards are supervised when s/he is in charge of them; calls for assistance before escorting a large group of wards. - Locks vehicle doors and uses appropriate restraints when transporting wards. - Conducts all required head counts and other security checks; logs supplies in and out as required. - Notifies Control immediately if a ward is discovered missing or escaping, or if facility keys or supplies are missing.	Takes extra precautions and care in maintaining security. For example: - Maintains close supervision of wards despite distractions or conditions encouraging relaxation of security (such as during a hospital visit when a physician says s/he can leave). - Conducts more frequent and thorough security checks than required. - Carefully questions visitors to assure their identity; detects unauthorized visitors who are trying to gain entry by impersonating a ward's relative. - In addition to notifying Control when security problems occur, immediately initiates appropriate follow-up (such as conducting a shakedown of the unit to look for missing keys, calling for a body count when an escape is rumored, or standing guard at a loose window until maintenance arrives).

The worker characteristic "Quantitative Skill" was renamed, "Arithmetic Skill" so that it more accurately reflected the type of abilities being assessed. Similarly, "Tolerance for Supervision and Structure" was renamed "Non-autonomy" and the worker characteristic dealing with vision and hearing was separated into two scales: one for vision and one for hearing.

Overall effectiveness. Finally, an overall effectiveness scale was constructed on which supervisors were asked to rate how well an incumbent met the standards and expectations of the job compared to other individuals with similar levels of experience. Ratings were made on the same 7-point scale ("Needs Improvement" to "Excellent") used for the behavioral summary scales.

The final criterion rating scales consisted of 9 scales developed from the critical incidents data, 27 scales based upon the worker characteristics, and a single, overall effectiveness scale.

CHAPTER FIVE: DATA COLLECTION

Once the experimental selection battery and criterion ratings materials had been developed and assembled, the tests were administered to a sample of juvenile counselor incumbents and job performance ratings collected from the incumbents' superiors. Following are the efforts undertaken to secure a sample of juvenile counselor incumbents of sufficient size and composition to permit an accurate and efficient assessment of the validity of the experimental selection battery.

Predictor Data Collection

Predictor data collection began with the setting of targeted sample sizes for each of the participating agencies. Sample sizes were based upon relative agency size; however, smaller agencies were requested to test a proportionally larger number of their incumbents than were larger agencies. The initial sampling plan called for testing 428 juvenile counselors from 21 probation departments.

Department coordinators were contacted by phone to secure permission to test the requested number of juvenile counselors. As agency participation was voluntary, a certain amount of negotiation was involved in arriving at an agreed-upon number of incumbents to be tested in each agency. In spite of the heavy and repeated requests for department resources imposed by this project, department coordinators were universally willing to go the "extra mile" to ensure that testing ran smoothly.

Information about the testing process was sent to each coordinator outlining the purpose of testing and reviewing the procedures to be involved. An important part of the information sent to the coordinators concerned the method by which incumbents were to be selected for testing. Because job performance ratings were to be collected for all incumbents tested, a selection method was devised that attempted to minimize the burden put upon each agency's supervisory staff. This sampling method called for the ordering of supervisors based upon their experience. Incumbents were then chosen by selecting all of the incumbents of the most experienced supervisor, all of the incumbents of the second most experienced supervisor, and so on until the targeted number of incumbents had been selected.

Testing was scheduled to occur during June 1984 but various organizational problems led to a number of delays so that testing was not completed until later that year. The first problem to occur was the withdrawal of two participating agencies, thereby reducing the targeted sample by 34 to 394. In order to offset this loss, PDI contacted other participating agencies in an attempt to secure additional participants but was unsuccessful in its effort. The additional step of contacting agencies that had participated during the job analysis phase but that were not involved in testing was undertaken. No additional incumbents were available from these agencies, however.

At that point, the maximum possible sample size for the validation phase stood at 394. However, it was not reasonable to assume that 394 usable test batteries would be completed. A more reasonable estimate, allowing for persons who failed to appear for testing, etc., would be that approximately 85% of the batteries would be completed. Thus, the expected sample size stood near 335. In order to secure additional subjects, STC consultants contacted agencies that had not previously participated in any aspect of the project. These agencies supplied an additional 5 incumbents, raising the total targeted sample to 399.

By October of 1984, phone canvassing of the participating agencies made it clear that no additional incumbents were available for testing. A count of the data collected to that point revealed a sample size of 189. In order to increase the number of incumbents tested, STC consultants contacted several agencies offering to reimburse them for the overtime payments they would incur as a result of testing additional incumbents ($75 per participant). Despite this offer, only 5 additional juvenile counselors were tested, bringing the total sample size to 194.

Criterion Data Collection

Soon after the experimental selection batteries were shipped to coordinators, job performance rating materials were also shipped in sufficient quantity to permit supervisors to rate the performance of all incumbents scheduled to be tested. These materials consisted of a performance appraisal booklet and two response booklets. The performance appraisal booklet introduced the performance rating task and provided general instructions about the rating process as well as detailed instructions for the completion of each rating scale. Each response booklet contained space to rate the performance of five incumbents. Each supervisor also received a cover memo providing more information about the ratings task (e.g., where it fit into the overall research effort). enlisting the supervisor's assistance in maintaining the confidentiality of the ratings he/she made, and directing the supervisor to the department coordinator in the event that questions about the ratings process arose.

Job performance ratings were returned by a total of 84 juvenile counselor supervisors from 17 agencies. These raters reported having been employed as supervisors for an average of 7.6 years. Fourteen of the supervisors were female, the rest (70) were male. Fifteen supervisors were black, six were Asian, six were Hispanic, and 54 were white (two supervisors did not report their ethnic status). About one-third of the supervisors reported that they worked in maximum security facilities with the remaining two-thirds fairly evenly distributed among minimum security, medium security, and "other" facilities. These 84 supervisors rated the performance of 183 incumbents on each of 9 behavioral summary scales, 27 worker characteristic scales, and the overall effectiveness scale representing a total of 6,771 ratings.

As predictor and criterion data arrived, checks were made to ensure that every juvenile counselor incumbent tested had also been rated. In those cases where performance ratings did not arrive for a tested incumbent, department coordinators were contacted and asked to follow-up with the appropriate supervisor to ensure that the ratings were made. Despite frequent requests to follow-up on ratings collection, it was not possible to secure job performance ratings for every juvenile counselor who completed the experimental selection battery.

Final Validation Sample

The criterion-related validation strategy employed in this study relied upon job performance ratings as the yardstick against which the validity of the predictors was to be judged. The final validation sample (i.e., those juvenile counselor incumbents whose data were analyzed to determine the validity of the predictors) consisted of those persons for whom <u>both</u> usable predictor and usable criterion data were collected. As was noted, 194 juvenile counselors were tested but only 183 had their job performance rated by their supervisor. Thus the maximum possible number of incumbents for whom both predictor and criterion data were available was 183. Unfortunately, there were other ways incumbent data were lost from the final validation sample. The missing data problems encountered in this project and the impact of those problems upon the size of the final validation sample are discussed in the following. Table 5.1 lists by county the number of incumbents tested and the number lost to the validation sample for the reasons enumerated. Following this discussion, the sample demographics are presented and the technical adequacy of the sample defended.

<u>Missing Name</u>. Incumbents' names were used to match predictor and ratings data. Incumbents who failed to put their names on their answer sheets prevented this match from being made for their data. Two cases were lost due to missing names.

Table 5.1

Composition of the Final Validation Sample

County	Number of JCs Tested	Lost in Validation Sample: Predictors w/o Name	JCs not Rated	Low Tenure	Not Well Known	Final N
Alameda	0	-	-	-	-	0
Conta Costa	12	-	-	-	-	12
El Dorado	0	-	-	-	-	0
Fresno	0	-	-	-	-	0
Humboldt	3	-	-	-	-	3
Kern	25	-	-	1	1	23
Los Angeles	32	-	-	2	3	27
Monterey	0	-	-	-	-	0
Nevada	3	-	-	1	-	2
Placer	6	-	-	-	-	6
Sacramento	19	2	5	1	-	13
San Bernadino	12	-	-	-	-	12
San Diego	19	-	1	1	1	16
San Francisco	0	-	-	-	-	0
San Joaquin	18	-	-	1	-	17
San Mateo	13	-	5	-	-	8
Santa Barbara	8	-	-	-	-	8
Sonoma	3	-	-	-	-	3
Stanislaus	9	-	-	-	-	9
Tulare	5	-	1	-	-	4
Ventura	1	-	-	-	-	1
Yuba	6	-	-	-	-	6
		Final Validation Sample				170

Missing Ratings. A second way in which cases were lost from the validation sample was through missing ratings data. Despite repeated efforts to follow-up with department coordinators, ratings were not obtained for 12 incumbents who had completed the experimental selection battery.

Low Tenure. The validation sample was intended to consist of only experienced incumbents (i.e., those on the job for at least six months). The job performance of incumbents with six or fewer months experience would be difficult to rate because supervisors may not have had an adequate opportunity to observe these incumbents performance under a variety of job conditions. In order to ensure that the job performance ratings of all incumbents accurately assessed their ability to perform the job, incumbents with six or fewer months tenure were not included in the final validation sample.

Inaccurate Ratings. When rating the performance of their incumbents, supervisors were also asked to indicate how well they knew the incumbent whose performance they were assessing. Incumbents whose ratings were made by supervisors who said that they did not know that incumbent "well at all" (i.e., rating of 4) were dropped from all analyses. In order to judge the validity of the predictors, an incumbent's true level of performance must be known. Obviously, the ratings of an incumbent not at all well known by his/her supervisor could not be an accurate measure of the incumbent's true job performance.

Demographics of Final Validation Sample

As a result of the data losses noted, the final validation sample consisted of the 170 juvenile counselor incumbents for whom accurate performance ratings were available. The demographic composition of the final sample is detailed in Tables 5.2 - 5.4. The sample is largely male and approximately one-third of the juvenile counselors are members of protected groups (principally, blacks or Hispanics). Most of the juvenile counselors in the final sample have some form of post high school education (i.e. Associate Degree, B.S., or B.A. Degree) and more than one-half of the sample have been juvenile counselors for two or more years (only 11% have less than one year's experience on the job).

Table 5.2

Gender and Ethnic Status of Final Validation Sample[a]

Group	Number	Percentage
Gender		
Male	116	68
Female	54	32
Ethnic Status		
Black	38	23
Native American	1	0
Asian	5	3
Hispanic	25	15
White	97	58
Other	2	1

[a]Totals may vary due to missing data.

Table 5.3

Education and Tenure of Final Validation Sample[a]

Group	Number	Percentage
Education		
High School	4	2
Tech./Vocational	7	4
Associate Degree	36	22
B.A./B.S. Degree	84	51
Some post-graduate	21	13
Post-graduate Degree	12	7
Tenure		
Less than 3 months	b	
3-6 months	b	
6 months to 1 year	12	7
1 to 2 years	33	20
2 to 5 years	56	33
More than 5 years	68	40

[a]Totals may vary due to missing data.

[b]JCs with fewer than 6 months tenure were not included in the final validation sample.

Table 5.4

Relationship of Incumbents to Supervisors in Final Validation Sample[a]

Group	Number	Percentage
Length of Time Supervised		
Less than 3 months	6	4
3 to 6 months	11	7
6 months to 1 year	46	27
1 to 2 years	51	30
2 to 5 years	45	27
More than 5 years	10	6
How Well Known by Supervisor		
Very well	101	62
Pretty well	52	32
Not too well	10	6
Not at all well	b	-

[a]Totals may vary due to missing data.

[b]Incumbents known "not at all well" were dropped from the final validation sample.

Technical Adequacy of Sample

One concern in conducting a criterion-related validation study is whether enough incumbents can be tested to permit a technically adequate evaluation of the validity and fairness of the predictors. Several analyses were performed to determine the technical adequacy of the sample collected in this study (See "Report on the Technical Adequacy of the Final Validation Sample for Juvenile Counselors" for details, available from the Board of Corrections). The major issue was whether the sample size provided sufficient statistical power to detect predictor validity and subgroup differences should they have existed. The analyses uniformly demonstrated the adequacy of the validation sample for all of the analyses performed during the course of the study. Thus, the 170 juvenile counselor incumbents for whom predictor and job performance data were available provided a more than sufficient basis for judging the validity of those predictors as measures of juvenile counselors' job performance.

The analyses reported in the following chapters were based upon all available data. This ensured that the results of those analyses were as stable as possible. As a consequence, the sample size varied slightly from analysis to analysis. This occurred because an incumbent who was missing data essential to one analysis might have had all the data required for another. Thus, the sample size would have varied because the incumbent was excluded from the first analysis but included in the second. Exact sample sizes for the analyses are reported in the tables summarizing the analyses. Slight variations in sample size are to be expected for the reason given above.

CHAPTER SIX: ANALYSES AND RESULTS

Many different analyses of the predictors and criteria were conducted to document their validity and fairness. The goal of all analyses was to obtain a combination of predictors that simultaneously:

1) would validly predict job performance;

2) was fair (i.e., did not unfairly impact on protected groups); and

3) was administratively practical for all agencies participating in the STC Program to implement.

The analyses conducted to meet the above goals can be divided into three categories: criterion analyses, predictor analyses, and validation and fairness analyses.

Criterion analyses included those necessary to determine: a) that the job performance ratings were reliable; b) how best to combine the 37 rating scales into a single measure of overall job performance; and c) whether there were specific aspects of performance differentially predictable by one or more of the 15 tests.

Predictor analyses included those necessary for final revision of the tests (especially Reading and Observational Memory), to: a) document the psychometric quality of the tests; and b) describe the level of performance obtained by the juvenile counselors in the final validation sample.

Validation and fairness analyses were those necessary to determine: a) how best to combine the tests from the experimental selection battery into a useful, valid predictor of job performance; and b) that the selected tests did not unfairly discriminate against candidates from protected groups.

While each of the above categories of analysis are conceptually distinct, there was considerable overlapping of analyses. Each category will be described separately although occasional reference to analyses from other categories will be necessary.

Criterion Analysis

The criterion analyses had as their basic goal the development of a reliable measure of overall job performance. This criterion would serve an important role in later validation analyses. An additional and equally important goal of the criterion analyses was to discover whether there were specific aspects of job performance that could be better predicted by single tests than by the Juvenile Counselor Examination. The presence of such differentially predictable aspects of job performance would have broadened the usefulness of the validation research as it would have resulted in a "menu" of selection instruments from which agencies could pick according to their needs.

Data collection and cleaning. Incumbents' job performance was rated by their supervisors on 9 behavioral summary scales, 27 worker characteristic scales, and on a single, overall effectiveness scale (see Chapter Four). To preserve the confidentiality of the performance ratings, supervisors were provided pre-addressed envelopes for mailing their rating materials directly to PDI. No one other than PDI staff members and the supervisor making the ratings had access to the performance ratings of any individual juvenile counselor.

When the ratings arrived at PDI, they were reviewed for completeness and assigned an identification number. Ratings were then keypunched. The keypunchers identified any ratings where there was a question as to the intended rating. PDI staff reviewed all ratings so identified and determined the correct rating value or set that value to "blank", indicating a missing value.

The mean rating value for each of the 37 individual rating scales is shown in Table 6.1. As can be seen, juvenile counselors varied in the ratings they received on the 37 scales. For example, performance in the area of Reading is more highly rated than is performance in the area of Social Skills. It should be noted that comparisons of this sort are only appropriately made between ratings from similar scale types (i.e. worker characteristic to worker characteristic or behavioral summary scale to behavioral summary scale) and are valid only for average group performance. To compare an individual's ratings across scales, it would first be necessary to convert the ratings into standardized units. This is necessary because a rating of 4 on a scale with an average of 4.5 represents poorer relative performance than does a rating of 3 on a scale with an average of 2.5.

The overall performance of juvenile counselors in the final validation sample ranges from "needs improvement" to "excellent". The mean rating for the overall effectiveness scale was 4.65 with a standard deviation of 1.23 indicating that 90% of all juvenile counselors received ratings between 3 and 7 on the 7-point scale.

Table 6.1
Descriptive Statistics for 37 Job Performance Ratings Scales

Scale	Number of JCs Rated	Avg.[a]	Stnd. Dev.
Behavioral Summary Scales			
Maintaining Security	183	4.81	1.27
Searching	183	4.72	1.23
Alertness to Potential Problems	183	4.95	1.33
Record Keeping/Report Writing	183	4.51	1.41
Handling Emergencies/Crises	178	4.73	1.37
Dealing with Emotionally Distressed Wards	183	4.69	1.41
Day-to-Day Interactions with Wards	183	4.81	1.35
Interacting with Other Staff and Parents of Wards	183	4.62	1.22
Commitment to Job	182	4.71	1.47
Worker Characteristic Scales			
Department-Specific Knowledge	180	4.68	1.34
Knowledge of California Juvenile Justice	183	4.10	1.26
Knowledge of Psychology	183	4.52	1.30
Street Knowledge	183	4.63	1.33
Safety/First Aid	183	4.89	1.06
Operating Communications Equipment	183	4.84	1.18
Oral Communications	183	4.70	1.30
Reading	183	4.91	1.24
Written Communications	183	4.65	1.30
Arithmetic Skill	178	4.75	1.11
Skill in Observing Behavior	183	5.01	1.20
Driving	173	4.66	1.07
Observing and Searching	183	5.01	1.13
Self-Defense/Physical Restraint of Others	183	4.60	1.35
Organizing/Supervising Group Activities	183	4.67	1.41
Social Skills	183	4.63	1.39
Motivation/Dependability	183	4.73	1.50
Stress Tolerance	183	4.53	1.29
Dominance/Self-Esteem	183	4.61	1.33
Tolerance of Supervision and Structure	183	4.74	1.26
Flexibility	179	4.41	1.25
Tolerance of Negative Aspects of the Job	183	4.34	1.23
Investigative Ability	181	4.41	1.24
Personal Organization	183	4.56	1.18
Ability to Meet Visual Requirements	177	4.86	1.18
Ability to Meet Hearing Requirements	177	4.89	1.19
Ability to Meet Physical Requirements	182	4.84	1.42
Overall Job Performance	183	4.65	1.23

[a]Ratings made on a 7-point scale.

The distribution of scores is somewhat negatively skewed, indicating that more scores are clustered toward the higher ratings than would be expected in a normal distribution.

Inspection of the intercorrelations of the 37 ratings scales indicated that ratings within categories were not more highly interrelated than were ratings across categories. For example, the average intercorrelation of the 9 behavioral summary scales is .49 and the average intercorrelation of the 27 worker characteristic scales is .45. The average correlation of the 9 behavioral summary scales with the 27 worker characteristic scales is .44 indicating that neither of these rating formats contributed significant unique method variance to the ratings data. This finding allows the ratings to be combined across all scales as is discussed in the next section.

Developing an overall criterion. The principal goal of the criterion analyses was to develop a single measure of job performance that would serve as the criterion for subsequent validation analyses. A criterion-related research design requires that there be a single criterion against which the validity of the overall selection battery is judged.

Inspection of the data in Table 6.1 indicated no reason to give precedence to any one of the individual rating scales as a measure of overall effectiveness. By itself, the 7-point overall effectiveness scale was not appropriate as the principal criterion for two reasons. First, individual scales do not provide as reliable measures as do combinations of scales, nor do they provide sufficient variation in criterion scores to permit an accurate assessment of the magnitude of the relationship between predictor and criterion. Second, the overall effectiveness scale may have meant different things to different raters. For example, rater A may have placed more weight on Oral Communications while rater B may have weighted Vision and Hearing more heavily. Thus, incumbents might not have been rated on the same basis from rater to rater. Because the overall effectiveness scale was inappropriate and because no other scale stood out as deserving special attention, an overall effectiveness score was constructed by averaging across all 37 rating scales.

The overall effectiveness criterion (hereafter called overall criterion) was computed by: a) standardizing each rating scale (T-scores; mean = 50.00, SD = 10.0); b) computing an average rating for each scale type (i.e., behavioral summary scales, worker characteristic scales, and overall rating); and c) taking the average rating across the three averages. As long as an incumbent received ratings on 90% or more of the 37 scales, missing ratings were treated by averaging across those scales for which ratings were made. Standardizing the scales prior to computing the overall criterion resulted in each type scale contributing equally to the final average. The above method resulted in an overall criterion with a mean of 50.4 and a standard deviation of 7.3 (based upon 170 cases). Table 6.2 presents the correlation of this

overall criterion with each of the individual rating scales. The average of the 9 behavioral summary scales correlated .94 with the overall criterion, and the average of the 27 worker characteristic scales correlated .89 with the same criterion. The magnitude of these correlations is due in part to the fact that the behavioral summary and worker characteristic scales each comprise one-third of the overall criterion. Thus, the correlations are inflated by the fact that they, in large part, represent the relationship of an average with itself. The average of the behavioral summary scales correlated .76 with the average of the worker characteristic scales. The average of the behavioral summary scales and the average of the worker characteristic scales correlated .84 and .76, respectively, with the overall effectiveness scale.

An internal consistency approach was taken to compute the reliability of the overall criterion. Criterion reliability was found to be .94. Originally it was intended that multiple ratings would have been collected for each incumbent in the validation sample. However, discussions with department coordinators indicated that it would be very difficult to locate even two supervisors who were sufficiently familiar with an incumbent's performance to provide accurate ratings data. The internal consistency approach used is the same that was described in Chapter 3 to estimate the reliability of other ratings data. This approach is somewhat susceptible to overestimating the ratings reliability due to a tendency on the part of raters to form a global impression of incumbents' performance and not to distinguish between different levels of performance on the different ratings scales (halo error). The rating instructions warned raters against this tendency and requested that care be taken in making the ratings. Nevertheless, the reliability reported above will slightly overestimate the true reliability of the ratings.

Ratings reliability is important in validation research only because it places a limit on the validity of the predictors. The less reliable the ratings, the lower the maximum possible validity. Reliability estimates are commonly used to correct validity coefficients for criterion unreliability. Such corrections always increase the magnitude of the estimated validity, with the degree of increase being inversely proportional to the magnitude of the reliability estimate. Because the method used to estimate the reliability of the ratings slightly overestimates that reliability, it is a conservative estimate. That is, its use to correct the validity estimate for criterion unreliability will not cause a large increase in the magnitude of the validity coefficient.

Table 6.2

Correlation of 37 Job Performance
Ratings Scales with the Overall Criterion

Scale	Corr. w/overall criterion[a]
Behavioral Summary Scales	
Maintaining Security	.76
Searching	.66
Alertness to Potential Problems	.76
Record Keeping/Report Writing	.64
Handling Emergencies/Crises	.76
Dealing with Emotionally Distressed Wards	.75
Day-to-Day Interactions with Wards	.72
Interacting with Other Staff and Parents of Wards	.70
Commitment to Job	.74
Worker Characteristic Scales	
Department-Specific Knowledge	.66
Knowledge of California Juvenile Justice	.55
Knowledge of Psychology	.63
Street Knowledge	.54
Safety/First Aid	.52
Operating Communications Equipment	.46
Oral Communications	.66
Reading	.52
Written Communications	.55
Arithmetic Skill	..47
Skill in Observing Behavior	.78
Driving	.39
Observing and Searching	.72
Self-Defense/Physical Restraint of Others	.50
Organizing/Supervising Group Activities	.69
Social Skills	.77
Motivation/Dependability	.75
Stress Tolerance	.70
Dominance/Self-Esteem	.72
Tolerance of Supervision and Structure	.66
Flexibility	.68
Tolerance of Negative Aspects of the Job	.64
Investigative Ability	.61
Personal Organization	.63
Ability to Meet Visual Requirements	.50
Ability to Meet Hearing Requirements	.51
Ability to Meet Physical Requirements	.47
Overall Job Performance	.96

[a]Number of cases varies from 177-183; all r's are significant at alpha = .05

Search for additional criteria. Apart from the development of an overall criterion, the criterion analyses had as an additional goal the development of criteria relating to more specific aspects of job performance. To the extent that such specific criteria were predictable using one or more of the tests in the experimental selection battery, agencies needing to assess those specific areas of job performance would have valid instruments at their disposal.

The search for additional criteria was conducted via factor analysis of the ratings data. The ratings intercorrelation matrix was divided into three submatrices: one containing the intercorrelations of the behavioral summary scales, the second containing the intercorrelations of the worker characteristic scales, and a third containing the intercorrelations of the behavioral summary and the worker characteristic scales. None of the submatrices contained the overall effectiveness rating scale. An iterative, principal components analysis, followed by varimax rotation to simple structure, was performed for each submatrix as is further described.

Analysis of each submatrix began with the extraction of a single factor. This was followed by additional analyses extracting two, three, . . . N, factors at each step. After each additional factor was extracted, the factor loading matrix was examined to determine whether the resulting solution made sense in light of what was known about the juvenile counselor job. Additionally, the traditional factor solution criteria of eigenvalue magnitude and variance accounted for were monitored. The iterative procedure continued until the extraction of the next factor either made no sense, or drastically exceeded typical rules of thumb for acceptable factor solutions (e.g., eigenvalue less than 1.0).

A two-factor solution was determined to be most appropriate for the behavioral summary scales submatrix and a five-factor solution was accepted for the worker characteristics submatrix. As these two solutions were judged to have considerable overlap, the factor analysis involving both behavioral summary and worker characteristic scales was taken as the most appropriate basis on which to construct additional, specific criteria. Descriptions of the resulting factors are given in Table 6.3. Figure 6.1 indicates how each of the 36 behavioral summary or worker characteristic scales loaded on the five-factor solution judged to be most appropriate. It should be noted that the "all or nothing" loadings shown in Figure 6.1 represent the forced simple structure solution that was used to compute scores for each additional criterion.

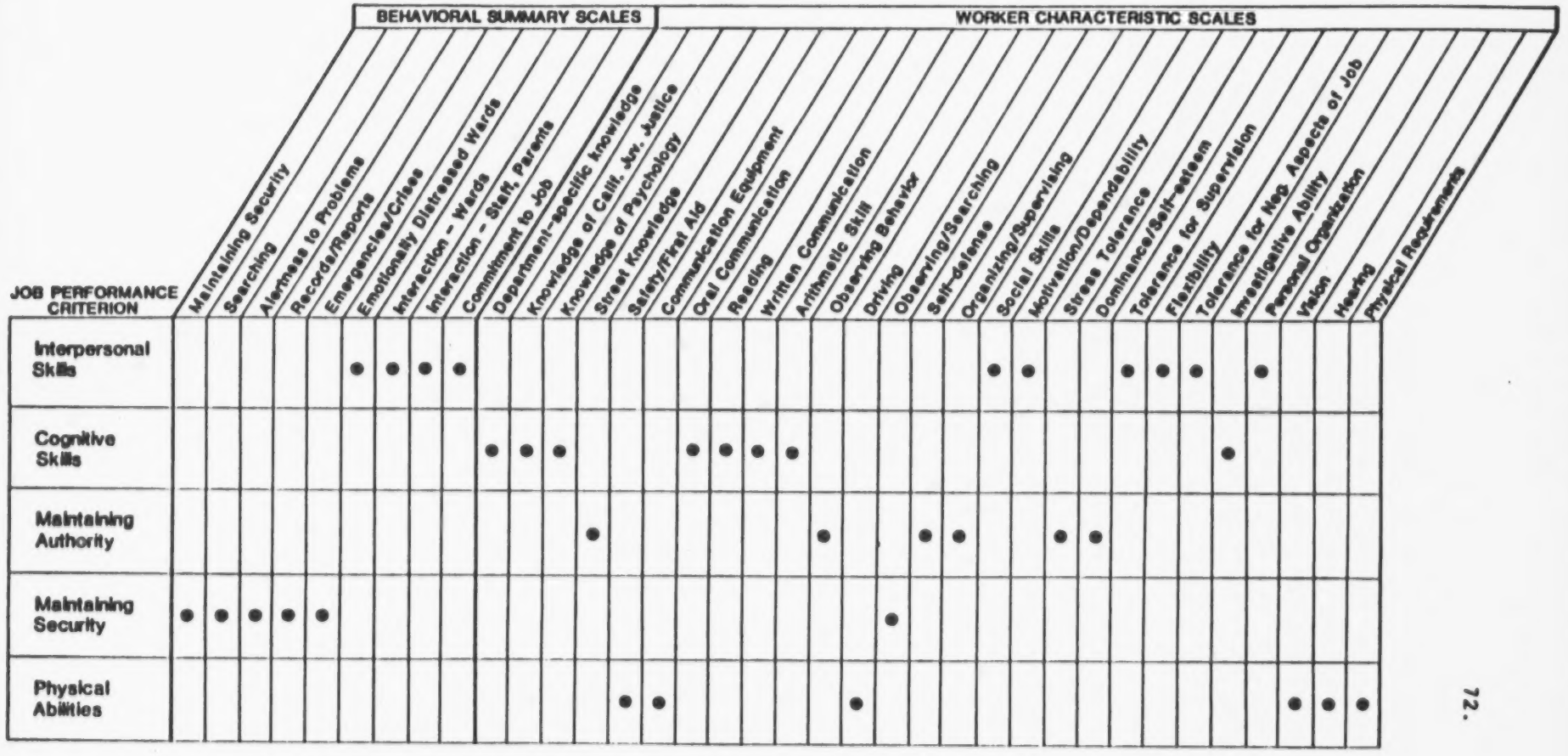
Criterion scores were determined by averaging across the standard scores (T-scores) of the rating scales loading on a particular factor. This method of computing factor scores differs from more commonly used methods in that in the latter, factor loadings are used to weight the contribution of each variable to each factor score. In such an approach every variable contributes to a factor score -- even those

TABLE 6.3

Factor Descriptions of
Five Specific Job Performance Criteria

Criterion	Description
Interpersonal Skills	This criterion involves skills necessary for harmonious relationships with others. These skills involve flexibility, tolerance, and motivation, as well as skills in interacting with wards and other staff members.
Cognitive Skills	This criterion revolves around the 3 R's -- reading, writing, and arithmetic. It involves the ability to speak and write well, and requires general knowledge about psychology and the California Juvenile Justice System.
Maintaining Authority	This criterion assesses a counselor's ability to maintain a position of authority. It includes the ability to organize and supervise group activities, to handle stressful situations, and to observe behavior. A successful counselor will be proficient in self-defense techniques and street knowledge and will also have high self-esteem.
Maintaining Security	This criterion assesses incumbents' ability to remain alert to potential problems, to deal with emergencies, to keep accurate counts and records, and to be skilled in conducting searches.
Physical Abilities	This criterion requires good vision and hearing as well as skill in searching and the ability to meet the physical requirements of the job.

Figure 6.1 Simple Structure Factor Loadings of 36 Ratings Scales on Five Specific Job Performance Criteria



● Indicates that the rating scale was used to define the job performance criterion and contributes to an incumbent's score for that criterion.

variables not considered to be part of a particular factor. The approach used in this project was to assign a rating scale only to the factor on which it loaded most clearly. Factor scores were the unit weighted average of all ratings that define a given factor. Due to the nature of factor analysis, these factor scores, although derived from an orthogonal (varimax) rotation, are not uncorrelated.

To summarize, the difference between typical methods of computing factor scores and the method used in this project is that the typical method yields factor scores that are uncorrelated but that are computed using variables not included in the factor definition. The current method yields factor scores based only upon those variables that define each factor, but which are substantially intercorrelated. Table 6.4 presents the descriptive statistics for each of the five additional criteria as well as the intercorrelations of the five factor scores.

Criterion descriptive statistics. The criterion analyses yielded a total of six criteria: a criterion of overall effectiveness and five criteria of the more specific aspects of job performance (Interpersonal Skills, Cognitive Skills, Maintaining Authority, Maintaining Security, and Physical Abilities). Table 6.5 presents descriptive statistics for these criteria broken out by incumbent gender and ethnic status (the latter only for those groups represented in numbers sufficient to permit accurate calculations). Female juvenile counselors received lower ratings than male counselors on two of the specific criteria (Maintaining Authority and Physical Abilities). They received higher ratings on Interpersonal Skills, and essentially the same ratings on Cognitive Skills, Maintaining Security, and the overall criterion. Minority counselors were rated higher than non-minority counselors on Maintaining Authority and were rated essentially the same as non-minorities on Physical Abilities. They were, however, rated lower than non-minorities on Interpersonal Skills, Cognitive Skills, Maintaining Security, and the overall criterion. There was concern that the lower performance ratings of minority juvenile counselors on some of these criteria might have been due to bias by the raters rather than due to true differences in the average performance. This concern led to the calculation of the average criterion scores by the ethnic status of raters as well as by the ethnic status of incumbents. Data for gender are given in Table 6.6. Care must be taken not to overinterpret these data as there are very few ratings available from female supervisors.

Table 6.4

Descriptive Statistics and Intercorrelations
for Five Factor Score Criteria

Criterion	Avg.	S.D.	Correlation with: 1	2	3	4	5
1. Interpersonal Skills	50.2	7.6	--				
2. Cognitive Skills	50.2	7.5	.63	--			
3. Physical Abilities	50.1	7.7	.70	.73	--		
4. Maintaining Security	50.5	7.6	.72	.63	.72	--	
5. Practical Knowledge	50.1	7.6	.54	.59	.66	.52	--

Table 6.5

Descriptive Statistics for the Overall and Specific Job Performance Criteria

Group		Overall	Specific Criteria[a] 1	2	3	4	5
Gender							
Males (116)	Mean	50.4	49.9	50.3	50.6	50.5	50.5
	SD	7.6	7.7	7.6	8.0	8.0	7.8
Females (54)	Mean	50.3	50.8	50.0	49.3	50.5	49.3
	SD	6.6	7.3	7.1	7.2	6.7	7.2
Ethnic Group							
Blacks (38)	Mean	49.9	49.8	49.4	51.4	49.6	50.5
	SD	7.6	8.4	8.3	7.7	8.3	7.9
Hispanics (25)	Mean	48.9	48.4	48.2	49.8	49.8	48.5
	SD	7.2	7.1	5.8	7.4	6.3	7.0
Whites (97)	Mean	50.5	50.4	50.4	49.2	50.6	50.2
	SD	7.3	7.4	7.5	7.7	7.8	7.6

[a] 1 - Interpersonal Skills

2 - Cognitive Skills

3 - Maintaining Authority

4 - Maintaining Security

5 - Physical Abilities

Table 6.6

Average Overall Criterion Scores by Supervisor and Incumbent Gender

Incumbent		Supervisor: Male	Supervisor: Female
Male			
	N	106	19
	Mean	49.7	52.1
	S.D.	8.3	5.8
Female			
	N	39	18
	Mean	49.1	50.7
	S.D.	6.3	7.9

Table 6.7

Average Overall Criterion Scores by
Supervisor and Incumbent Ethnic Status

Incumbent	Supervisor		
	Black	Hispanic	White
Black			
N	10	3	26
Mean	50.1	57.4	47.0
S.D.	7.4	11.4	6.8
Hispanic			
N	5	---	20
Mean	50.6	---	48.9
S.D.	5.6	---	7.5
White			
N	6	7	90
Mean	49.8	49.2	50.1
S.D.	4.5	7.9	7.8

Ratings data for ethnic groups are given in Table 6.7. Care must be taken not to place too much emphasis on the sparse data in this table. There were no statistically significant differences in the average ratings made by raters from different ethnic groups. Black and Hispanic supervisors gave higher ratings than did white supervisors, but there was no significant statistical interaction between supervisor and incumbent ethnic groups. Taken together, the data in Tables 6.6 and 6.7 give no indication that the mean differences in criterion scores of various subgroups are due to anything besides true differences in the ability of members of those groups to perform the juvenile counselor job tasks.

Predictor Analysis

Predictor analyses were conducted to ensure that all predictor data were accurately collected and scored, to evaluate the 15 tests in terms of their psychometric quality, and to describe the performance of the validation sample on each of the tests.

Data collection and cleaning. Incumbents recorded their test answers on machine readable answer sheets. Upon receipt of the answer sheets, the first step in predictor analysis was to check each sheet to ensure that all marks were dark enough for the optical scanner to "see" and to make sure that the ID number on every answer sheet corresponded to the number printed on the test booklet.

Once checked, the answer sheets were read by an optical mark reader and incumbents' responses were transferred to the computer where they would be further scored and analyzed. As part of scanning, any answer sheet that contained one or more apparently missing responses was flagged for inspection. This was necessary to ensure that marks too light for the scanner to "see" were not missed. Where necessary, marks were darkened and the answer sheets rescanned.

Scale revisions and final psychometrics. Once all predictor data had been accurately scanned, the 15 tests comprising the experimental selection battery were scored and examined for indications of needed revisions. The goal of scale revision was to achieve a final set of reliable tests for consideration during the validation analyses.

Scale revisions began with the computation of item-total correlations and item difficulty indices. Item-total correlations were computed using the point-biserial correlation and item difficulty was computed as the proportion of incumbents passing each item. Both statistics were computed for every item and distractor. The item statistics were computed using data from all 194 completed experimental selection

batteries. Items for which the keyed alternative correlated negatively with total scores were dropped from the tests. Additionally, typographic errors during the assembly of the experimental selection battery resulted in one item being dropped from each of the Reading, Arithmetic, Vocabulary, and Verbal Reasoning tests.

Item difficulty statistics for the Verbal Reasoning test indicated that few incumbents were successfully completing the last five items on that test and so the items were dropped.

Psychometrics for all tests (after revisions) are given in Table 6.8. Reliability coefficients (internal consistency KR-20) are presented for the cognitive ability tests; published estimates are given for the noncognitive tests. As can be seen, the tests are all sufficiently reliable to be used in pre-employment screening. In addition, the tests were all of moderate difficulty, indicating that they were appropriate for use in the candidate population.

Descriptive statistics. Table 6.9 gives the mean (average score) and standard deviation of each test in the experimental selection battery. Table 6.10 shows the intercorrelations among these 15 tests. Tables 6.11 and 6.12 give the descriptive statistics of each test broken out by incumbents' gender and ethnic status, respectively. Female juvenile counselors performed better on all of the cognitive ability tests than did their male counterparts. On the noncognitive tests, however, the male counselors generally scored higher than the female counselors. The exceptions to this trend included the Sociability, Control, and Responsibility tests, where females scored slightly higher.

Black and Hispanic juvenile counselors received scores that were considerably below those of white juvenile counselors for all of the cognitive ability tests and for most of the noncognitive tests. Minority counselors received higher scores than non-minority counselors only for the Sociability and the Stress Reaction tests.

Table 6.8

Final Item Statistics for the
Experimental Selection Battery

Test	Number of Items Original	Dropped	Avg. Disc.	Avg. Diff.	rel.[a]
Observational Memory	30	--	.15	.77	.56
Reading	70	24	.28	.80	.79
Arithmetic	35	1	.33	.76	.85
Language Usage	35	--	.26	.71	.76
Vocabulary	35	1	.33	.78	.83
Analogies	35	--	.30	.66	.81
Verbal Reasoning	30	6	.34	.61	.80
Sociability	35	2	.22	.68	.70-.85
Empathy	29	6	.12	.54	.84
Control	24	--	.34	.77	.84
Socialization	49	6	.19	.73	.65-.80
Responsibility	36	1	.19	.66	.65-.85
Stress Reaction	24	--	.48	.79	.90
Dominance	44	5	.26	.66	.65-.80
COIB	58	--	.10	.65	.70

[a]Reliability for noncognitive tests are from published sources; the remainder are KR-20 estimates.

Table 6.9

Descriptive Statistics on 15 Individual Tests for 194 Juvenile Counselors

Test	N[a]	Avg.	Stnd. Dev.
Observational Memory	184	22.71	3.47
Reading	194	53.95	8.64
Arithmetic	194	26.88	5.00
Language Usage	194	25.79	4.39
Vocabulary	194	27.70	4.30
Analogies	194	23.45	6.04
Verbal Reasoning	186	15.05	5.07
Sociability	192	22.45	4.88
Empathy	192	13.71	3.23
Control	191	17.72	4.30
Socialization	191	30.75	4.27
Responsibility	192	22.52	4.55
Stress Reaction	191	17.29	5.70
Dominance	191	25.87	5.52
COIB	184	29.71	3.67

[a]Ns less than 194 are due to tests not being scored for individuals with excessive missing responses; see text for details.

Table 6.10

Intercorrelation of 15 Predictor Tests

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Obs. Memory	-														
2. Reading	.25*	-													
3. Arithmetic	.27*	.43*	-												
4. Language Usage	.13*	.44*	.53*	-											
5. Vocabulary	.20*	.55*	.36*	.56*	-										
6. Analogies	.20*	.60*	.50*	.57*	.59*	-									
7. Reasoning	.19*	.59*	.37*	.42*	.45*	.58*	-								
8. Sociability	.17*	.11	.09	.04	.17*	.13*	.18*	-							
9. Empathy	.17*	.16*	.02	.06	.19*	.11	.11	.61*	-						
10. Control	-.09	-.14*	.01	.02	-.02	-.05	-.11	-.05	-.08	-					
11. Socialization	-.01	-.08	.08	.04	.03	-.01	-.01	.27*	.27*	.29*	-				
12. Responsibility	.09	.10	.07	.22*	.28*	.17*	.10	.48*	.43*	.26*	.37*	-			
13. Stress Reaction	.00	-.01	.00	-.09	.08	-.04	-.10	.48*	.37*	.32*	.38*	.50*	-		
14. Dominance	.10	.11	.08	.12	.16*	.12	.11	.67*	.60*	.04	.27*	.54*	.46*	-	
15. COIB	.01	.14*	.03	-.06	.07	.03	.11	.13*	.22*	.01	.32*	.16*	.11	.18*	-

*Statistically reliable correlation at alpha = .05.

Table 6.11

Descriptive Statistics on Individual Tests
by Incumbent Gender

	Males[a]		Females[b]			
	Avg.	S.D.	Avg.	S.D.	t[d]	ES[c]
Observational Memory	22.6	3.5	23.0	3.4	.72	.12
Reading	53.2	9.0	55.5	7.7	1.66	.27
Arithmetic	26.8	4.9	27.1	5.3	.42	.06
Language Usage	25.4	4.4	26.4	4.4	1.44	.23
Vocabulary	27.4	4.5	27.9	3.9	.75	.12
Analogies	23.1	5.8	24.1	6.3	1.08	.17
Verbal Reasoning	14.5	5.0	15.7	4.9	1.47	.24
Sociability	22.1	5.3	22.6	4.1	.67	.10
Empathy	13.8	3.4	13.3	2.8	-.82	-.16
Control	17.6	4.3	17.8	4.3	.30	.05
Socialization	30.7	4.2	30.5	4.7	0.37	-.05
Responsibility	22.4	4.9	22.6	4.0	.30	.04
Stress Reaction	17.4	6.0	16.6	5.4	-.87	-.14
Dominance	26.0	5.8	25.3	5.2	-.71	-.12
COIB	30.1	3.8	29.1	3.4	-1.67	-.27

[a]N = 125

[b]N = 58

[c]ES = Effect size (standardized mean difference); negative values indicate that females scored, on average, below males.

[d]None of the differences were statistically significant at alpha = .05

Table 6.12

Description Statistics on 15 Individual Tests
by Incumbent Ethnic Status

Test	Blacks[a] Avg.	S.D.	Hispanics[b] Avg.	S.D.	Whites[c] Avg.	S.D.	F	ES
Obs. Memory	22.4	3.5	21.8	3.8	23.1	3.5	.89	-.
Reading	48.8	9.4	50.5	9.2	56.8	6.5	8.37[e]	-.
Arithmetic	24.0	4.4	26.9	4.3	27.7	5.1	3.95[e]	-.
Language Usage	22.8	4.4	25.4	4.7	26.8	3.9	5.93[e]	-.
Vocabulary	25.0	4.9	26.1	4.3	29.0	3.2	7.21[e]	-.
Analogies	19.6	6.2	20.8	5.7	25.4	5.0	8.20[e]	-.
Verbal Reasoning	11.1	5.3	13.0	4.9	16.5	4.2	8.83[e]	-1.
Sociability	23.9	3.9	21.2	5.6	22.2	4.9	2.89[e]	.
Empathy	13.8	3.0	13.1	3.6	13.8	3.2	.93	
Control	18.8	3.4	17.9	4.1	17.2	4.6	.86	-
Socialization	30.8	4.1	29.9	4.1	30.8	4.5	0.64	-.
Responsibility	22.9	4.4	21.7	5.5	22.6	4.5	.58	-.
Stress Reaction	18.3	5.6	17.8	6.1	16.5	5.8	1.06	.
Dominance	26.9	4.5	23.6	6.1	26.1	5.7	2.23	-.
COIB	29.4	3.8	28.7	4.0	30.1	3.6	.90	-.

[a] N = 40

[b] N = 28

[c] N = 105

[d] Standardized mean difference contrasting blacks and Hispanics to whites; negative value indicat lower average scores by blacks and Hispanics.

[e] Statistically significant at alpha = .05

VALIDITY AND FAIRNESS

Validity and fairness analyses had as their basic goal the evaluation of the potential of each test to provide valid, fair predictions of juvenile counselor job performance. Valid tests are those for which high scores indicate a higher probability of successful job performance than do low scores. Fair tests are those for which no average score differences occur among gender or ethnic groups except where differences of a similar magnitude occur for the criterion.

Once the validity and fairness of each individual test was determined, a combination of the tests could be chosen to form the Juvenile Counselor Examination, thereby meeting the project goal of developing and validating entry-level selection procedures for the juvenile counselor position.

An additional goal of the validity and fairness analyses was to determine whether any of the five specific job performance criteria were differentially predictable in a fair and valid manner, either by individual tests or by a different combination of tests than that chosen to predict overall job performance.

Adverse impact and test fairness. Tests for which two or more groups receive scores that are statistically different on average are said to discriminate between the groups. Used in this limited sense, the term discrimination simply indicates that there is a statistically reliable difference between the groups (i.e., a difference likely to be found every time the groups are tested). The presence of group differences for a selection procedure does not automatically disqualify that procedure however. The Uniform Guidelines on Employee Selection (1978) discuss the relationship between two important concepts related to the fair use of selection procedures. The first of these is adverse impact and the second is test fairness.

Adverse impact (as defined by the 4/5ths rule of thumb) exists whenever the selection ratio for members of a protected group is 4/5ths (80%) or less than that of the group with the largest selection ratio. Unfairness is defined in the guidelines as occurring when members of one group obtain lower scores on a selection procedure and when similar differences are not found for measures of job performance. The Guidelines also note that test fairness must be investigated (if technically feasible) whenever a selection procedure results in adverse impact, but not when adverse impact is absent.

From the perspective of selection procedures, discrimination in the limited sense described above can be either fair and therefore permissible, or unfair and therefore not permissible. In evaluating a selection procedure the first step then is to determine whether adverse impact exists. Adverse impact is a joint function of the magnitude of

group differences and the overall selection ratio used during selection. Even very large group differences may not result in adverse impact if the selection ratio is high enough.

Many different models have been offered for use in distinguishing fair from unfair discrimination. One model (Cleary, 1968) holds that group differences in predictor scores can only be considered fair if differences of the same magnitude occur for the criterion measure. In statistical terms, this translates into a test of the similarity of regression equations computed separately for each subgroup involved in the fairness analysis. Selection decisions for any valid test or procedure depend upon the use of a single regression equation to predict the expected performance of all candidates. A selection procedure is unfair to the members of any group whose performance is systematically underpredicted by a common regression equation.

The standardized mean difference between groups (effect size) is an important statistic in this determination of test fairness. In the Cleary model, tests biased against minority groups (large negative effect size) will only be fair if a difference of comparable magnitude exists for criterion scores. From the perspective of adverse impact, selection procedures with large negative effect sizes require that very low passing scores be set in order that sufficient numbers of protected group members be selected. For either model, effect sizes indicating bias against protected groups warn of the necessity of further analyses to determine whether adverse impact is present, and, if adverse impact exists, whether the test is fair to protected groups.

Validity and fairness of the individual tests. Following all necessary revisions, each of the tests in the experimental selection battery was correlated with the overall criterion. At the same time, average scores and associated descriptive statistics were computed for all members of the final validation sample as well as separately by incumbent gender and ethnic group. The descriptive statistics were reported in Tables 6.11 and 6.12. Table 6.13 presents the validity coefficients for the 15 tests in the selection battery. As can be seen in this table, for two tests (Reading and Arithmetic) the validity estimates were statistically significant at the .05 level. Of these two, Arithmetic stands out as the best single predictor of job performance when magnitude of the validity coefficient is the sole criterion of "best predictor". None of the other tests demonstrated any promise of providing accurate predictions of juvenile counselor job performance.

Tables 6.11 and 6.12 presented descriptive statistics by gender and ethnic status as well as fairness statistics, respectively for each test. For each, means and standard deviations by subgroup are given,

Table 6.13

Correlations of 15 Individual Tests
with Overall Job Performance

Test	Corr. w/job perf.	p[a]
Observational Memory	.08	.15
Reading	.14	.03
Arithmetic	.23	.00
Language Usage	.09	.13
Vocabulary	.02	.40
Analogies	.10	.10
Verbal Reasoning	.05	.29
Sociability	.03	.34
Empathy	.01	.47
Control	.01	.46
Socialization	.02	.39
Responsibility	.06	.20
Stress Reaction	.04	.29
Dominance	.07	.18
COIB	.10	.10

[a]Probability that a correlation this large or larger would be obtained due to chance. Probabilities below .05 are considered statistically significant.

along with the appropriate test for determining whether differences between groups are statistically significant. Finally, an index of the magnitude of the difference between groups is reported. This index, effect size, gauges the distance between group means relative to the pooled variation in scores. Conceptually, effect size estimates are interpreted in the same way as are z scores. For example, an effect size of .30 indicates that the mean of one group lies .30 pooled standard deviations above the mean of the second group and, therefore, that 62% of the members of the first group receive scores greater than the average score of the second group. In groups with equal means, 50% of persons in either group receive scores above the mean of the other group.

The data in Table 6.11 show that neither the Reading test nor the Arithmetic test had statistically significant mean gender differences. Thus, those tests shown to be valid predictors of job performance are also fair predictors for males and females.

Fairness data for the three ethnic groups with large enough samples to permit calculation of reliable statistics are given in Table 6.12. These statistics are quite similar to those reported for gender except that a one-way analysis of variance rather than a t test was used to determine whether the differences among groups were statistically significant. The effect size measures for ethnic groups contrast the combined group of black and Hispanic counselors to white counselors.

The two tests for which statistically reliable validity estimates were found (i.e., Reading and Arithmetic) also displayed statistically reliable mean differences across the three ethnic groups. In both cases, these differences favored whites. Comparison with the data in Table 6.5 shows that the mean differences among ethnic groups on each of the tests are much larger than the corresponding mean differences on the overall criterion. Thus, for ethnic groups, the tests showing promise as valid measures of job performance are not fair predictors for the three ethnic groups.

Developing a final selection battery. Following the evaluation of the fairness and validity of each of the 15 individual tests, it was necessary to determine which combination of those tests would be recommended for use as the Juvenile Counselor Examination. Several constraints influenced the choice of tests to be included in the final battery. Principal among these was the necessity that the battery be valid and either a) incur no adverse impact, or b) provide fair predictions of the performance of candidates from all groups. Of considerable additional importance was that the final selection procedures assess candidates' potential in all important job areas, that the final set of tests require no more than 2 1/2 hours to administer, and that scoring be as simple as possible so as not to compromise the operational validity of the final battery.

A core set of tests was selected from those that made up the experimental selection battery according to the following procedure. First, only those two tests (Reading and Arithmetic) that were significantly correlated with the overall criterion were considered candidates for inclusion as entire, intact tests. The Arithmetic test was thus included in the final selection battery in its entirety. Subsequently, each of the six passages in the Reading test was separately analyzed with respect to its validity and standardized ethnic group difference. The three passages with the greatest group differences were deleted from the final Reading test. Thus, the core set of tests included the original Arithmetic test and an abbreviated Reading test. A total score was then computed for each juvenile counselor incumbent by summing his/her scores on the two selected tests. This combination of tests was found to be a valid predictor (r with overall criterion = .26, p < .05), but resulted in a sizeable mean difference favoring whites (effect size = -.62), suggesting that adverse impact would be present.

Several procedures are available that permit the use of apparently unfair tests in a manner that does not result in adverse impact against protected groups. For example, separate selection lists can be constructed and candidates can be selected from both lists in numbers that do not violate the 4/5ths rule of thumb. Another method is to adjust test scores by adding a constant value to some scores, thus removing the mean differences among groups. As neither of these methods were administratively feasible for the Board of Corrections, the following method was used to expand the final selection battery while maintaining or increasing its validity and decreasing the magnitude of the mean differences among ethnic groups.

Because none of the remaining tests from the experimental selection battery could be used in its entirety without increasing the bias against minority groups, a search was undertaken at the item level. Because the cognitive ability tests were highly intercorrelated and because they were the principal source of mean differences favoring whites, the pool of items to be searched was narrowed to those items from the noncognitive tests.

The final validation sample was randomly split into halves and the effect size of each item (blacks and Hispanics versus whites) was computed for every item in both samples. Items with positive item validities and that favored minorities in <u>both</u> samples were selected for use in the final battery.

In reviewing these statistics, it was noted that several items from two noncognitive tests (Sociability and Empathy) would favor minorities in both random halves if the keying of those items was reversed. The content of each of these items was examined to ensure that reversing their keys would not result in rewarding candidates for endorsing undesirable traits or behaviors. As long as reversing an item key was judged not to cause such problems, the keys were reversed and the items added to those already selected for the battery.

This procedure identified 32 items drawn from a maximum possible 279 items from the noncognitive tests. Table 6.14 indicates the source of the items selected using the methods described above. Predictor scores were recomputed for all incumbents by summing across the items that comprise the Juvenile Counselor Examination. These scores were found to correlate .33 with the overall criterion, indicating that the battery was a valid measure of job performance.

Although every precaution was taken to minimize its occurrence, the method used to select additional items for the final selection battery had some probability of having capitalized on chance variation among items. The effect of such capitalization would be to slightly overestimate the validity of the final predictor scores. Thus, the true validity of the final selection battery lies somewhere in the interval from .26 (validity prior to the addition of the 32 items) to .33 (validity after the addition of the 32 items).

Fairness data for the final battery are given in Table 6.15. As can be seen, none of the group differences are statistically significant. The effect size for gender on the predictor is .16, while the corresponding effect size on the overall criterion is -.14. For ethnic groups, the predictor-score effect size is -.15, while the criterion effect size is -.01. That the mean group differences are not statistically significant indicates that they are no larger than would be expected due to chance differences that occur between any groups of similar size. Nevertheless, the predictor effect size for ethnic groups is larger than the corresponding criterion effect size. This indicates the need to examine whether the Juvenile Counselor Examination results in adverse impact against minority candidates.

As discussed previously, effect size is one of two critical factors in determining whether a test will violate the 4/5ths rule of thumb; the other factor being the cutoff score chosen. Score distributions for minority and non-minority incumbents were checked for normality. Scores for both groups were determined to be normal. In addition, the variances of the two score distribution were found to be statistically equivalent. These analyses demonstrated that minority and non-minority scores distributions differed only in terms of their means. This finding permitted the computation of the expected adverse impact that would result from various overall selection ratios.

Table 6.14

Source of Items Selected to
Augment the Final Selection Battery

Source	Items Selected[a]
Sociability[b]	9
Empathy[c]	7
Control	4
Responsibility	6
Socialization	2
Dominance	3
Stress Reaction	3
COIB	2

[a]Number selected sums to 36 due to four items which are scored on two scales.

[b]Two items are reverse keyed.

[c]Four items are reverse keyed.

Table 6.15

Gender and Ethnic Group Statistics
for Juvenile Counselor Examination Scores

Group	N	Mean	S.D.	F[a]	ES
Males	125	77.6	8.4		
Females	58	78.0	8.3	1.03	.16
Blacks	40	77.1	8.1		
Hispanics	28	78.2	8.9		
Whites	105	78.9	8.5	.976	-.15

[a]Neither of these F ratios is statistically significant at alpha = .05.

These computations showed that as long as the overall selection ratio was greater than .20 (raw score of 85), no violation of the 4/5ths rule occurs. A selection ratio of .20 is more stringent (only 20% of all applicants pass) than is recommended for an examination of this type. Even if a cutoff score of 85 were set, use of the final selection battery to "select" candidates from the final validation sample, the Juvenile Counselor Examination would not have resulted in adverse impact against minority group members. Assuming that the effect size for ethnic groups remains near -.15 in applicant groups and further assuming that minority and non-minority groups have approximately equal score distributions, in order to avoid adverse impact selection ratios must be kept above .20, a value above that recommended for use with this battery (See Chapter 7). As no adverse impact was found for minority juvenile counselors, no further analyses were necessary.

Adverse impact analyses for females revealed that there were no overall selection ratios that would lead to adverse impact. Because no adverse impact against female juvenile counselors was found, no further analyses were necessary.

Differential predictability of specific criteria. Correlations of the 15 tests and the final predictor with five specific criteria of job performance are shown in Table 6.16. As can be seen, there were several individual tests whose use resulted in valid predictions of one or more of the specific criteria. As can also be seen, the final predictor was nearly as valid or more valid a predictor as were any of the individual tests. Tables 6.11 and 6.12 report the likely adverse impact of the individual tests against female and minority candidates. Because the individual tests invariably result in considerable adverse impact, none of them would be both fair and more valid than would the Juvenile Counselor Examination. It was therefore concluded that the specific criteria were most predictable using the Juvenile Counselor Examination scores, so no additional predictors were investigated.

Table 6.16

Correlations of Individual and Final
Predictors with the Five Factor Criterion Measures

Predictor	Factor Criterion[b] 1	2	3	4	5
Observational Memory	-01	10	02	07	-07
Reading	08	09	03	12	-02
Arithmetic	17[a]	26[a]	12	27[a]	03
Language Usage	08	12	-10	06	-13[a]
Vocabulary	02	14[a]	-11	-02	-15[a]
Analogies	10	11	-07	04	-07
Verbal Reasoning	00	13[a]	-09	-01	-06
Sociability	-01	14[a]	11	-03	10
Empathy	-04	05	07	-05	04
Control	10	00	-07	03	-08
Socialization	08	11	03	-06	07
Responsibility	08	15[a]	07	00	07
Stress Reaction	-02	-02	03	-01	-03
Dominance	-01	14[a]	12	04	06
COIB	10	00	12	07	06
Final Predictor	28[a]	26[a]	24[a]	31[a]	08

Note. Decimal points omitted.

[a]Statistically significant at alpha = .05

[b]1 = Interpersonal Skills
2 = Cognitive Skills
3 = Maintaining Authority
4 = Maintaining Security
5 = Physical Abilities

CHAPTER SEVEN: DISCUSSION AND RECOMMENDATIONS

The Juvenile Counselor Examination is a fair and valid indicator of potential to successfully perform the juvenile counselor job. PDI therefore recommends that the Board of Corrections adopt this selection adopt this selection instrument (or some other equally valid and fair measure of the same abilities and personal characteristics) as part of the entry-level selection procedures for agencies participating in the STC program. The Juvenile Counselor Examination is not intended to serve as the sole or even as the primary basis for selection decisions. Its judicious use, in combination with other sources of information about candidates' potential (e.g., application appraisals, interviews, educational achievement, etc.), and according to the constraints set forth below will, however, increase the probability that selected candidates will be successful juvenile counselors.

The remainder of this chapter provides a description of the format and content of the Juvenile Counselor Examination and sets forth the more important constraints governing the use of the examination as one source of information about candidates' potential. Recommendations are made concerning: a) scoring of the examination; b) the data collection and documentation necessary to supplement the norms collected in this project; and c) and the efforts necessary to document the continued validity and fairness of the examination.

Exam Format and Content

The Juvenile Counselor Examination consists of a single test booklet containing items assessing candidates' cognitive and noncognitive knowledges, skills, abilities, and personal characteristics. The examination contains 117 items and has a total testing time, including time for instructions, of approximately 1 1/2 hours.

Two cognitive abilities are measured in Parts 1 and 2. The first of these (Reading) requires that candidates read three brief (1-3 paragraphs) passages and answer 10 to 14 questions about each passage. The questions are mostly factual and require only the ability to remember what was read (or to quickly review information). A few of the questions require that simple conclusions be drawn based upon information presented in each passage. Taken together, the Reading items measure the ability to read, understand, and recall written information. The passages are written at approximately a twelfth-grade level (Flesch index) and are intended to be similar in content to the types of information juvenile counselors read on the job.

Part 2 contains 35 items that measure candidates' facility with the arithmetic operations of addition, subtraction, multiplication, and division. Candidates are to apply these operations to simple multiple choice problems involving whole numbers, fractions, decimals, and percentages.

In Part 3 the Juvenile Counselor Examination contains 48 items that assess several personal characteristics found (in the job analysis) to be important components of successful juvenile counselor performance. These items were drawn from scales of the California Psychological Inventory, the Multidimensional Personality Questionnaire, and the Correctional Officers' Interest Blank (COIB). Following are the personal characteristics measured by the items. These descriptions borrow heavily from descriptions of the scales where the items originally appeared.

Approximately one-third of the items were drawn from scales (MPQ Control and CPI Socialization and Responsibility) that reflect a tendency to take a cautious, planful approach to life. Persons who receive high scores on these scales are described by others as being rule-respecting, self-controlled, responsible, and preferring to plan their activities rather than play things by ear. Another personal characteristic measured is dominance (CPI Do). Dominant persons are those who behave in an ascendant manner, who are self-confident, who can take the initiative and exercise leadership in a self-assured manner. Items were also selected from the MPQ Stress Reaction scale. High scorers are those people who are able to put fears and worries out of their minds easily and who are not troubled by emotional turmoil or feelings of guilt.

Approximately half of the items were drawn from the CPI Sociability and Empathy scales; as noted previously, approximately one-third of these items have had their keys reversed. In general, high-scoring candidates are those who are outgoing and participative and who are patient and affiliative by nature.

Care should be taken not to overinterpret the scores on the Juvenile Counselor Examination. Individual items were selected for inclusion on the basis of their relationship to the overall criterion (i.e., their criterion-related validity) and their effect on the mean group differences in scores. They were not selected to yield meaningful interpretations regarding underlying psychological constructs. Thus, the utility of these items lies in how well they predict, as a set, the overall criterion scores of juvenile counselor incumbents. Table 7.1 contains descriptions of the noncognitive scales from which these items were selected.

Table 7.1

Description of the Noncognitive Tests
Contributing Items to the Juvenile Counselor Examination

Test	Description
CPI Dominance	A subscale of the California Psychological Inventory that identifies individuals who behave in a dominant, ascendant manner and who, in interpersonal situations, can take the initiative and exercise leadership in a forceful, self-confident manner.
CPI Sociability	A subscale of the California Psychological Inventory that indicates a person's tendencies to be outgoing and participative and to seek out and enjoy social contact without being shallow or superficial.
CPI Responsibility	A subscale of the California Psychological Inventory measuring a person's tendency to believe that life is best governed by reason and to behave in a responsible, self-disciplined manner.
CPI Socialization	A subscale of the California Psychological Inventory measuring a person's tendency to be socialized, rule-respecting, prudent, and responsive to what others feel and think.
MPQ Stress Response	A subscale of the Multidimensional Personality Questionnaire that indicates persons who are able to put fears and worries out of their mind easily; quickly get over upsetting experiences; and who are not troubled by emotional turmoil or feelings of guilt.
MPQ Control	A subscale of the Multidimensional Personality Questionnaire indicating a person's tendency to be reflective, cautious, careful, rational, and sensible, and who likes to anticipate and plan activities and events.
COIB	The Correctional Officers' Interest Blank is a 40-item questionnaire that "tends to identify individuals with strong internal values and who can be counted on to behave in a dependable and rule-observing way."

Table 7.1 (continued)

Test	Description
Empathy	A series of items drawn from the California Psychological Inventory that identifies persons who are described as being "characterized by a patient and forebearing nature, by affiliative and socially ascendant tendencies and by liberal and humanistic political and religious attitudes." (Hogan, 1969).

The Juvenile Counselor Examination is intended to be a power rather than speed measure. The time limits suggested for the exam are such that most candidates will have ample opportunity to complete all items. The exam equally weights the cognitive and noncognitive items and is scored by counting the number of items that a candidate answered correctly.

Using the Juvenile Counselor Examination

The Juvenile Counselor Examination is intended to be used only as described in the User's Manual. This manual provides general information concerning the physical settings recommended for testing as well as verbatim proctor s instructions to be followed every time the exam is administered. A Candidate Information Booklet has been prepared that describes the purpose and nature of the exam. No candidate should be tested without having received the Candidate Information Booklet in time to review the information it contains.

The Juvenile Counselor Examination is a fair and valid indicator of candidates' potential to perform the juvenile counselor job but only if it is administered and scored according to the guidelines in the User's Manual. The validity and fairness of any other use of the exam cannot be guaranteed. Persons wishing to use the exam in a manner other than that described in the User's Manual must document for themselves the validity of the exam for that use. All items in the Juvenile Counselor Examination are copyrighted and permission to use the exam in part or in whole must be obtained from the Board of Corrections, Consulting Psychologists Press and the University of Minnesota Press.

Exam Security

Security of the test keys is of the utmost importance. If the security of the keys is breached, the validity of the entire examination is compromised. The governing principle of key security is: The fewer people to have access to the keys the better. The best possible arrangement would be to have all tests scored by the Board of Corrections and to allow only one or two persons access to the test keys. If central test scoring is not possible, the next best arrangement would be to allow only one person from each agency access to test keys. This person would be responsible to the Board of Corrections for the maintenance of test security.

Test keys should not be reproduced indiscriminately. All copies of the keys should be numbered and an accurate record should be kept of their location.

Security of test keys also depends upon the security of the exam booklets. Test booklets should be numbered and accurate records should be kept by the Board of Corrections as to which agencies have which booklets.

Periodic (semi-annual) inventories should be made to ensure that all testing materials can be accounted for and remain secure. The Board of Corrections must be informed immediately if examination materials are lost or destroyed so that the appropriate action can be taken.

Scoring and Data Collection

Detailed instructions for scoring the Juvenile Counselor Examination are contained in a separate confidential report. Briefly, scoring is accomplished via a set of scoring templates. Candidates will record their responses on an answer sheet. To score the examination, the appropriate scoring template is placed over the answer sheet and the number of responses visible through the template are counted. Parts 1-3 will be scored separately and overall scores will be formed by summing the Part 1 - Part 3 subscores. Space will be provided on the answer sheets to record a candidate's scores. Examination answer sheets are designed to permit easy transfer to an electronic scoring system.

Table 7.2 contains the frequency distribution of incumbent examination scores. These data show the number and proportion of candidates who received various final scores. Also shown are the actual and expected percentile equivalents of the scores. The actual percentile equivalents are simply the proportion of incumbents who received a score at or below a given value. Expected percentile equivalents reflect the proportion of incumbents who would score at or below a given score based upon a normal distribution of test scores with the same mean and standard deviation as were obtained in the incumbent sample. Because these data are based on the scores of incumbents, they may slightly overestimate the proportions of applicants expected to obtain any given score.

The data in Table 7.2 are the norms that PDI recommends be used to guide test scoring and interpretation (as discussed below) for initial administrations of the Juvenile Counselor Examination. It is important that the Board of Corrections record (at a minimum) the race, gender, and final score of every applicant tested with the Juvenile Counselor Examination. These data are essential for evaluating the appropriateness of the recommended minimum acceptable score and for monitoring the examination's validity and fairness. PDI recommends that the Board of Corrections also retain copies of all answer sheets so that item level analyses of the Juvenile Counselor Examination can be undertaken as necessary. If examinations are scored by individual agencies (not recommended), copies of all testing materials should be sent to the Board of Corrections. Agencies should retain copies of all testing materials for a minimum of one year, as these materials are likely to be subpoenaed should the examination be challenged in court.

The Board of Corrections is strongly encouraged to conduct annual audits of the Juvenile Counselor Examination to, at a minimum, check scoring accuracy, evaluate the current recommended minimum acceptable score, and evaluate gender and ethnic differences in test scores.

Table 7.2

Juvenile Counselor Examination Score Distribution

Raw Score	N	Percentile Equivalent	
		Actual	Expected
94	194	99	99
92	189	97	95
89	177	91	90
87	165	85	85
86	158	81	80
84	145	75	75
83	134	69	70
82	126	65	65
81	118	61	60
80	105	59	55
79	90	50	50
77	80	41	45
76	71	37	40
75	64	33	35
74	57	29	30
73	54	28	25
72	48	25	20
70	34	18	15
68	25	13	10
65	16	8	5
59	4	2	1

It is also recommended that the examination be revalidated against a job performance criterion after it has been operational for two years and every three to five years thereafter unless the annual audits indicate otherwise.

Score Interpretation and Cutoff Scores

Methods used to set cutoff scores and a recommended minimum acceptable score are contained in a separate report available from the Board of Corrections. The probability that candidates scoring within several score ranges would be average or better performers was computed. The recommended minimum acceptable score is that score for which there is a 50-50 chance of being an average or better performer.

Although a minimum acceptable score has been recommended, no statewide cutoff was set or recommended. Choosing an appropriate cutoff score requires consideration of factors that will vary from agency to agency. This makes the establishment of a statewide cutoff impossible. The cutoff score report discusses the factors that influence the setting of cutoff scores and provides guidelines for agencies to follow in establishing their own cutoff score. This will allow individual agencies considerable flexibility in determining how best to use the Juvenile Counselor Examination.

References

American Psychological Association. (1974). Standards for Educational and Psychological Tests. Washington, DC: Author.

American Psychological Association, Division of Industrial-Organizational Psychology. (1980). Principles for the Validation and Use of Personnel Selection Procedures. (2nd ed.). Berkeley, CA: Author.

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Equal Employment Opportunity Commission. (1978). Uniform Guidelines on Employee Selection Procedures. Federal Register, 43, 38290-38315.

Gough, H.B. (1982). Manual for the Correctional Officers' Interest Blank. Palo Alto, CA: Consulting Psychologists Press.

Overall, J.E., & Klett, C.J. (1971). Applied Multivariate Analysis (Ch. 7). New York: McGraw-Hill.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 86, 420-428.

STANDARDS AND TRAINING
FOR CORRECTIONS PROGRAM

STC

BOARD OF CORRECTIONS

TECHNICAL ADEQUACY REPORT

JUVENILE COUNSELOR

Copyright 1987 California Board of Corrections

In preparing the Juvenile Counselor Examination for the California Board of Corrections, Personnel Decisions, Inc. has taken into account the following documents concerning selection testing:

- The American Psychological Association, Standards for Educational and Psychological Tests, (1974).
- The APA Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures, 2nd edition, (1980).
- The Equal Employment Opportunity Commission, Uniform Guidelines on Employee Selection Procedures, (1978).

Report On the Technical Adequacy Of the Final Validation Sample For Juvenile Counselors

The original project plan for the PDI-STC Juvenile Counselor Validation Project called for testing a minimum of 500 juvenile counselor incumbents. For reasons elaborated herein the actual number of juvenile counselors in the final validation sample ranged from 170 to 194 (depending upon the analysis being conducted). This report documents the data collection efforts resulting in the final validation sample in order to show: (1) that every effort was made to collect usable data from as many incumbents as could reasonably be tested; and (2) that the number of incumbents actually tested is technically adequate to permit the reliable computation of all necessary statistics and analyses.

The data collected in this criterion-related validation project consisted of scores on 15 measures of cognitive and noncognitive skills and abilities (predictors), and job performance ratings in 36 specific areas, plus a rating of overall effectiveness (criteria). For most analyses (e.g., validity, fairness) it was necessary that data for both predictor and criteria be available while other analyses (e.g., norms) required only that predictor data be available.

Predictor Data Collection

Data collection began with telephone canvassing of all agencies that had participated in the job analysis phase of the project. Agencies were contacted to determine their willingness to participate in the testing phase of the project and to get an approximate census of the number of eligible incumbents (i.e., those on the job for six or more months) from each agency. Target samples were set based upon relative agency size, but smaller agencies were generally requested to test a proportionally larger number of their incumbents than were larger agencies. The initial targeted sample called for testing 428 juvenile counselors from 21 probation departments.

Once targeted samples were set, participating departments were contacted by phone to reconfirm their participation in the testing phase of the project and to secure permission to test the targeted number of incumbents. During the follow-up phone calls, two agencies informed us that, due to delays in the scheduled testing date, they would be unable to participate in testing. The withdrawal of these agencies reduced the overall targeted sample from 428 to 394.

The remaining 19 agencies were contacted again to see if additional incumbents were available for testing. The agencies committed an additional 10 incumbents to the testing phase, thus raising the total targeted sample to 404.

Because the targeted sample was still well short of the 500 incumbents called for in the project plan, agencies that had participated in the job analysis phase of the project but were not participating during testing were contacted to see if they might be able to supply additional incumbents. No agencies were able to provide additional incumbents for testing.

To further raise the number of incumbents tested, STC consultants contacted a number of agencies that had not previously participated in any phase of the project. Due to the efforts of STC, one agency joined the validation phase of the project and supplied an additional 5 incumbents, raising the total targeted sample to 409.

As it appeared that no additional incumbents were available for testing, materials were distributed to all agencies which had not received them and testing commenced. Unfortunately, not all agencies were able to meet their targeted sample and only 46% of the total targeted sample (189 of 409) was actually tested. The main loss of incumbents occurred in one agency where the participation of incumbents in testing was voluntary. Despite agency solicitation of 800 volunteers, only 20% of the target number of incumbents were actually tested.

A final effort was made to obtain additional incumbents from all eligible agencies throughout California. This effort netted an additional five incumbents, raising the total number tested to 194.

Criterion Data Collection

The initial sampling plan called for the collection of job performance ratings for all juvenile counselors tested. Ratings materials were supplied to participating agencies in sufficient quantity to permit rating of all tested incumbents. Coordinators were instructed to select incumbents for testing in a manner designed to minimize the administrative burden of collecting the job performance ratings. Nevertheless, and despite repeated contacts with agencies, ratings data were collected for only 183 of the 194 juvenile counselors for whom predictor data were also available.

Due to test booklets being returned without incumbents' names, incumbents with fewer than six months experience being tested, and ratings made by supervisors who indicated that they did not know the rated incumbents "well at all", the final number of juvenile counselors for whom both predictor and criterion data were available was 170.

Technical Adequacy of the Sample

There are three analyses in particular for which the technical adequacy of the final validation sample must be shown. These are the validation analysis, fairness analysis, and the development of test norms. Each analysis involves one or more specific statistical tests. Every

statistical test involves the testing of a null hypothesis (e.g., no validity, or no difference between groups) versus some, usually unspecified, alternative hypothesis. If an estimate of the magnitude of the effect being tested can be specified (e.g., validity greater than .30, mean difference of .80 standard deviations), the power of the test to detect that effect can be determined. Statistical power is simply the probability (expressed as the number of chances in 100) that the test will reject the null hypothesis if the effect being tested is as large or larger than its estimated value.

The power of all statistical tests is a direct function of the amount of data used to compute the test statistic — the more data, the more powerful the test. A given sample is technically adequate for a given statistical test if it is of sufficient size to permit a reasonably powerful test of the effect. To determine the power of a statistical test some estimate of the expected magnitude of the effect must be made. The smaller the expected effect, the larger the sample needed to achieve a reasonably powerful statistical test. Statistical power is directly analogous to the power of a microscope where more power is needed to see very small objects than is needed to see relatively larger objects. The technical adequacy of the final validation sample is estimated in the following for the three main analyses of the project.

Validity analyses. The technical adequacy of the final validation sample for validity analyses depends upon the power of the statistical test of the significance of the correlation between predictor and criterion. The correlation coefficient is based upon data from those juvenile counselors for whom both predictor and criterion data are available. In order to determine the power of the statistical test for a sample of 170 some estimate of the expected validity is required. Boehm (1982) reports that the average validity coefficient from 176 studies spanning 20 years was .22. This is the estimate used to test the power of the validity analyses for this project.

Schmidt, Hunter, and Urry (1976) provide a formula to determine the sample size necessary to achieve a desired power for a specified validity coefficient. The problem here is only somewhat different than that posed by Schmidt et al. We want to know the power for a sample of 170 and an estimated validity of .22. The Schmidt et al formula can be converted for use in answering this question. Thus it can be shown (see Appendix) that if the true validity of the final predictor in this project is at least .22, then 81 times out of 100 it will be detected as being statistically significant using a sample of 170. On this basis it is clear that the final validation sample in this project is technically adequate for the purpose of validation analyses.

Fairness Analyses. The technical adequacy of the statistical tests involved in fairness analyses depends upon their ability to detect subgroup differences in average predictor and average criterion scores. As was true for validity analyses, some estimate of the magnitude of the expected effect is necessary to determine the power of the statistical tests. While a straightforward estimate was available for validity, no such easily obtainable estimate exists for gender or race differences.

The final predictor in this study consists of a combination of cognitive and noncognitive measures that has not been previously examined. For general measures of intelligence is not uncommon to see differences of approximately one standard deviation between minority and nonminority groups, while males and females typically do not differ. For the cognitive component of the examination we might expect to find rather large ethnic group differences favoring whites but no gender differences. The magnitude of typical group differences for the tests that make up the noncognitive portion of the examination is not known. In any case, no sufficiently precise estimate of the expected magnitude of group differences on the Juvenile Counselor Examination is possible.

Without a reasonable estimate of the expected magnitude of group differences no precise determination of statistical power can be made. Figures 1 and 2 show the power of the $\underline{t}$ test (a statistical test for group differences) plotted as a function of the standardized mean difference between groups. Figure 1 reports the power of the test to detect male-female differences and Figure 2 shows power for detecting ethnic group differences. In Figure 2, three power curves are shown. These are, from left to right, power to detect differences between grouped minority and nonminority incumbents, between black and white incumbents, and between Hispanic and white incumbents. Note: Although not shown as such, the power curves in Figures 1 and 2 are symmetric about the Y-axis.

The power curves shown in Figures 1 and 2 indicate that for the sample sizes in this project, the statistical test for detecting group differences has sufficient power to detect large group differences and reasonable power to detect moderate group differences (magnitude = .25-.45 standard deviations); but that the power for detecting small group differences (0.1 - .25 standard deviations) is low.

Fairness analyses depend less upon detecting statistically significant differences between groups than they do upon estimating the magnitude of subgroup differences on predictor and criterion measures. For this reason less powerful statistical tests can be tolerated for fairness analyses (as opposed to validity analyses where statistical significance is the primary concern), as long as the sample is of sufficient size to provide a stable estimate of subgroup means and standard deviations.

Figure 1. Statistical power of a t test with 181 df as a function of effect magnitude.

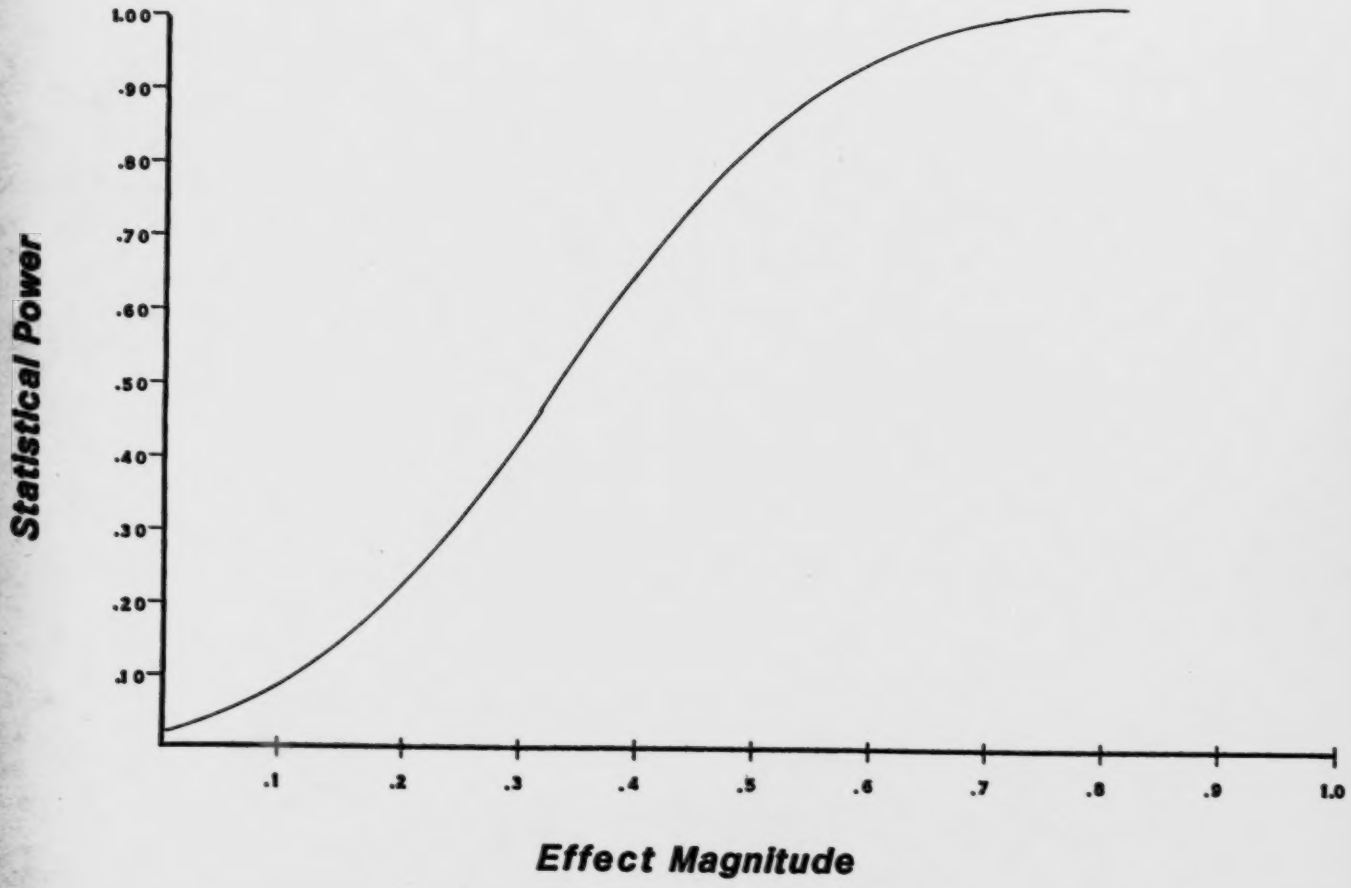
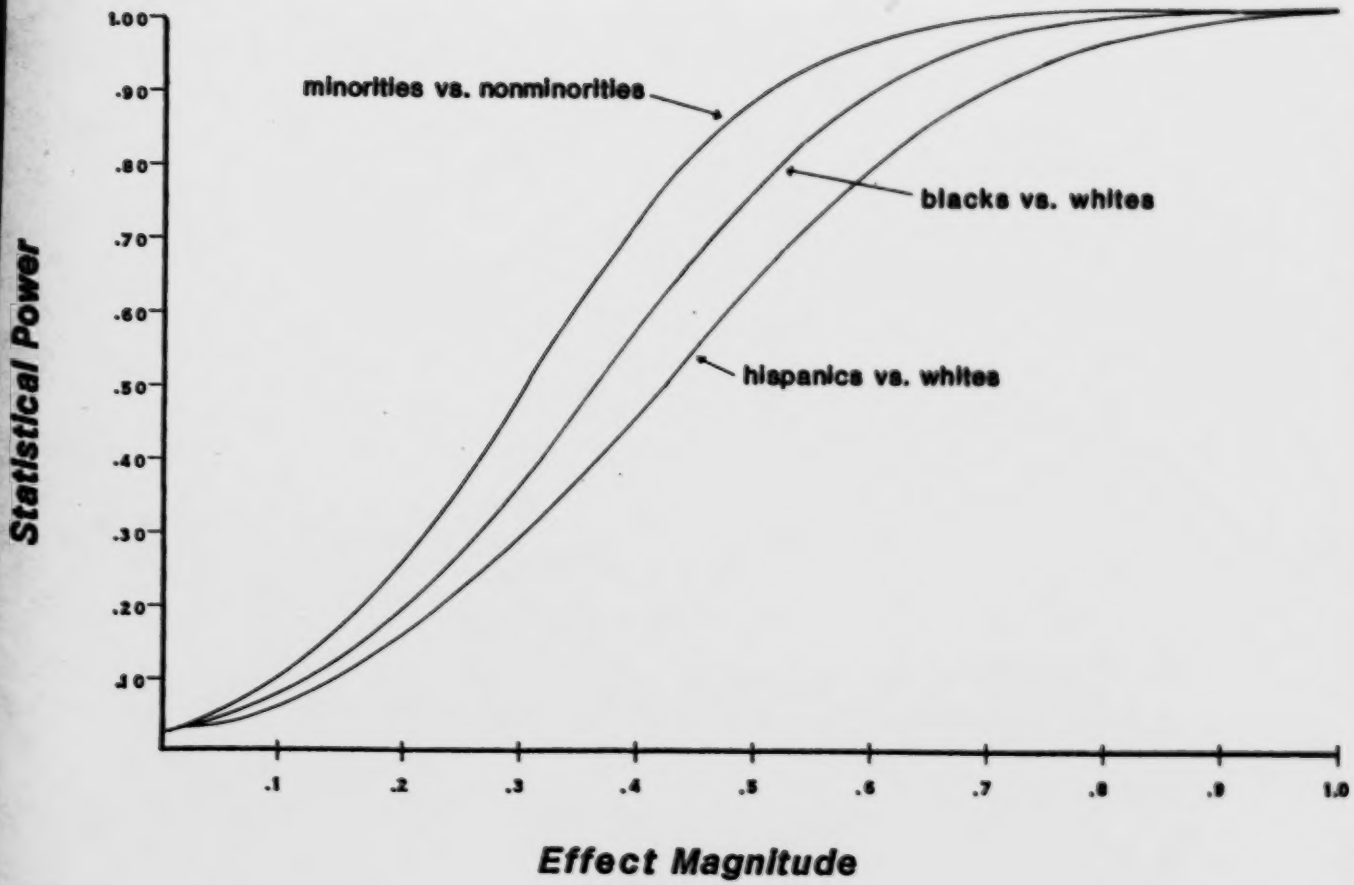


Figure 2. Statistical power of t tests with 171, 143, and 131 df as a function of effect magnitude.



The smallest subgroup in this study (Hispanics) has a sample size of 28. For the final predictor, a sample of 28 is sufficient to place a 95% confidence band of width 6.7 points about the group mean. In other words, we can be 95% confident that the true group mean for Hispanics lies within ± 3.3 points of the obtained mean score of 78.2. For the largest subgroup (whites) the corresponding 95% confidence interval spans 3.3 points. Thus, while the 95% confidence interval for the smallest subgroup is slightly more than twice that of the largest subgroup, it is small relative to the total range of scores (104 points) and therefore provides a quite accurate estimate of the true subgroup mean.

Estimating the adverse impact of the predictor upon subgroups (a major fairness analysis of this project) requires accurate estimates of subgroup mean scores. The sample sizes in this project are large enough to provide the needed accurate estimates and thus are technically adequate for the fairness analyses conducted.

Normative Data. Normative data must provide an accurate estimate of the performance of the validation sample on the Juvenile Counselor Examination. As was discussed, even the smallest subgroup in this project is large enough to provide quite accurate estimates of subgroup mean scores. Predictor data are available for a total of 194 juvenile counselor incumbents. Using these data we can be 95% confident that the true average score of juvenile counselors lies within ± 1.2 points of the obtained mean of 78.6. Thus there can be little doubt as to the expected performance of incumbents on the Juvenile Counselor Examination, thereby establishing the technical adequacy of the final validation sample for use in determining test norms.

Summary

The original project plan called for the testing of 500 juvenile counselor incumbents. Due to factors beyond the control of PDI or STC only 194 incumbents were tested. Of these 194 only 183 incumbents were rated as to their job performance. As a result, fairness and validation analyses were conducted using fewer data than was originally planned.

The adequacy of statistical tests for a given sample is a joint function of the sample size and the expected magnitude of the statistic (e.g., validity coefficient, group mean) being tested. The power of the three major analyses performed in developing and validating the Juvenile Counselor Examination has been described. In all cases, the statistical tests conducted in this project had sufficient power to detect effects of the magnitudes expected in this project. Thus the validation sample provides a firm basis for evaluating the appropriateness of using the Juvenile Counselor Examination to screen job candidates.

In the future STC should continue to collect gender, ethnic status, and examination score data in order to monitor the validity and fairness of the Juvenile Counselor Examination and to augment the norms upon which selection decisions will be made. Further recommendations concerning the use and monitoring of the Juvenile Counselor Examination can be found in the Juvenile Counselor Validation Report and related documents available from the Board of Corrections.

References

Boehm, V.R. Are we validating more but publishing less? Personnel Psychology, 1982, 35, 175-187.

Schmidt, F.L., Hunter, J.F., & Urry, V.W. Statistical Power in criterion-related validation studies. Journal of Applied Psychology, 1976, 61, 473-485.

Technical Appendix

Below is a brief outline of the computations used to determine the power of a sample of 170 to detect a correlation coefficient of .22 or greater.

Schmidt, Hunter, and Urry (1976) give the following equation for use in determining the sample size necessary to achieve an established test with specified power.

$$N = \frac{(Z_1 - Z_2)^2}{E(F_z)2} + 3.$$

Where: Z_1 = z score corresponding the alpha level of the statistical test (alpha is typically set at .05);

Z_2 = z score corresponding to the power of the statistical test;

N = The sample size necessary to achieve a given power for a specified alpha level; and

$E(F_z)$ = The expected value of the test statistic.

Solving the above equation to isolate Z_2 yields:

$$Z_2 = Z_1 - E(F_z) \sqrt{N-3}.$$

For juvenile counselors we want to know the power of a statistical test based upon N = 170 and $E(F_z)$ = .220 at an alpha level of .05 (Z_1 = 1.96). Substituting into this equation gives Z_2 = .88. Using a table of the normal distribution it is found that a z score of .88 corresponds to a power of .81.